Hybridization in *Candida* yeast pathogens

Verónica de Pinho Mixão

TESI DOCTORAL UPF 2020

Thesis supervisor:

Dr. Toni Gabaldón

Comparative Genomics Group

Until August 2019: Bioinformatics Department
Center for Genomic Regulation

From September 2019: Life Sciences Department
Barcelona Supercomputing Center

Universitat Pompeu Fabra Barcelona

Aos meus pais,
por estarem sempre lá para mim

“You must take your place in the circle of life. (…)
Remember who you are.”

Mufasa in The Lion King

Acknowledgements

When we apply for a job, we have to prove ourselves as the best person for the position. But what happens when we do not have much experience and specially when we want to change our field? Yes, that was my case in 2015. I wanted to leave the *wetlab* and enter in the world of bioinformatics, but my experience in the field did not pass from the use of online tools. So, there was no way I could sell myself as the best for a bioinformatics position, but I had to keep applying. It seems that it worked. In Christmas 2015, I came to an interview for such a position in Barcelona. I was so nervous on that day that I almost do not have any memory of it. Apparently, it went well. I got the job. This was my Christmas present. I will always be grateful to **Toni** for giving me this chance. A chance to change my life, to learn, and to work on something I really wanted. Toni, I want to thank you for this and more. Thanks for all the professional and personal advices you gave me during all these years, for guiding me in the right direction, and for being so understanding every time I made a mistake, or I had a personal drama. Thank you for everything!

I also want to thank all the members of the Comparative Genomics lab with whom I had the chance to work with. **Marina**, you are a special person, who has been always there for me since the rainy day in which you picked me in a bus stop. I am sorry for all the times I bothered you, but what can I say… "it is faster to ask Marina, than google", right? Thanks for everything you taught me. Thanks for making me go to the gym, for all the international dinners you hosted, all the advices you gave me, and for being always there to listen to me when I needed. **Cinta**, it was a pleasure to be your "left desk" for so many years. Thanks for being so nice and having always a smile. You deserve all the happiness of the world! **Hrant**, my OPATHY brother, *shnorhakalutsyun* for all the travels and all the PhD/OPATHY moments we shared, and of course for teaching me Armenian. **Miguel**, thank you for all your efforts to make this Portuguese girl a happier person, somehow you did it. **Irene** (and **Riccardo** of course), thank you for the dinners, escape rooms, and all the great moments we spent together. **Laia**, thank you for your constant happiness which made the days in the lab much better. **Ester**, **Susana**, and **Ewa**, you are great. Thank you for all the times you helped me and for being so nice. **Ernst**, thank you for being a

nice guy, even if you do not want us to know that. **Jesse**, after three years only seeing you in group meetings, it was very nice to share the space with you during these last months. I apologize for all the times that I "attacked" your mother tongue with all my mistakes. **Uciel**, thanks for solving our problems so many times. **Miki**, thanks for making me a more tolerant person, otherwise I would have sent a paper ball to your head every time you sing in the office. **Olfat**, thanks for bringing some fresh air to the lab. **Manu** thanks for helping with the computer. **Rosa**, thanks for all the nice discussions at lunch time. **Fran**, **Matteo**, **Marcello** and **Edoardo**, thank you for making my life inside and outside the lab so much better when you were there. **Leszek**, thanks for all the scripts you left in the lab… they saved my life! **Romina** and **Imma**, thanks for making our PhD lives much easier.

I also want to thank all the members of OPATHY network for this amazing journey, specially **Jonas** for keeping us organized and on track, and **Ahmed**, **Aimi**, **Antonio**, **Antonio**, **Elise**, **Frank**, **Giuseppe**, **Hrant** (again), **João**, **Marina** and **Mansoureh** for all the great moments we spent together. Thank you to the European Commission for funding this amazing Marie-Curie network, and my PhD.

To my **Friday night dinner's crew**, thank you for all the amazing moments we had. **Ombry**, thank you for everything!

And now it is time to change to Portuguese. **Hélder** e **Joana**, muito obrigada por terem sido a minha família em Inglaterra. Obrigada por me terem recebido tão bem e por tudo o que partilhámos, incluindo umas belas ostras em Norwich. **Mariana**, tão longe, mas sinto-te sempre tão perto, obrigada! **Sofia**, **Sofia**, **Bruno**, **Quina**, **Dora**, **Rita**, a vida levou-nos cada um para seu lugar, mas espero que nos volte a unir no futuro. E porque o caminho se faz caminhando, mas nunca sozinho, um agradecimento muito especial ao **Prof. Paulo Almeida** e à **Prof. Teresa Novo**, que me deram a oportunidade de aprender e evoluir, e sem os quais nunca chegaria aqui.

Às três estrelas mais brilhantes no céu: à **Nelinha** por todos os seus sorrisos e beijinhos, à **Avó Mixão** por todos os abraços e arrozes de cabidela de galinha (galo, não!) que deu a esta figurona, e ao **Avô Mixão** pelo seu carinho e por ter a capacidade ínfima de associar o

seu sentido de humor inteligente, a uma retidão de valores e pensamentos inigualáveis que o tornaram um dos homens mais fascinantes que conheci. Aos **Avós Pinho** por terem sempre um sorriso e um abraço pronto para me receber, e ainda me mimarem como se eu tivesse 10 anos. Ao melhor **Padrinho**, à **Tia Graça**, à **Pati** e à **Falela** por serem simplesmente espetaculares. Ao **Tio Nel**, à **Tia Nazaré**, à **Marta** e ao **Hugo** por estarem sempre disponíveis para um belo convívio e uma animação quando ia a casa. À **Micas**, ao **Carlos**, à **Daniela** e ao **Gabriel** por darem mais ritmo à vida. A todos eles, obrigada por me recordarem sempre como é bom estar perto de quem amamos. A toda a minha família que de uma forma ou de outra contribuíram para eu ser quem sou, muito obrigada!

Ao **nha cretcheu**, sem quem eu não estaria neste momento a escrever estes agradecimentos. Obrigada por mesmo longe teres estado sempre comigo nos bons e nos maus momentos. Amo-te.

Por fim, o agradecimento mais especial de todos é para **os meus pais**. Eles que sempre estiveram lá para mim quando precisei de celebrar ou de chorar. Eles para os quais não tenho palavras capazes de descrever toda a sua imensidão. Obrigada por serem simplesmente os melhores!

Moltes gràcies a tots!

Verónica de Pinho Mixão
Barcelona, 82nd day of confinement by COVID-19.

Abstract

Fungal infections are a growing problem for human health, causing ~1,350,000 deaths each year. *Candida* species are among the most important fungal pathogens. Although *Candida albicans* is the most common cause of *Candida* infections, many other *Candida* species have emerged as pathogens. How pathogenicity is evolutionary acquired is unknown, but previous studies point to a role of hybridization in its development. Hybrids are chimeric organisms that may present unique phenotypic features and higher plasticity, which possibly facilitates adaptation to new niches, such as humans.

This thesis studied the genomic features of *Candida* pathogens, with a special focus on hybrid species and their evolution. Specifically, it asked the questions of how spread are hybrids among *Candida* species, and what are the processes that drive the evolution of their genomes. To this end, genomes from 141 isolates belonging to 13 *Candida* species were analyzed and compared, to reconstruct their features and evolution. The hybrid nature of six of these species, including *C. albicans*, is reported here for the first time. These results suggest that hybrids are widespread among *Candida* species. The factors underlying this apparent high propensity for hybridization are yet to be understood. However, our results indicate that these hybrids only face mild genomic incompatibilities, and that drift dominates their genomic evolution. Overall, this thesis supports an important role of hybridization in the emergence of new yeast pathogens and provides novel insights on the evolutionary aftermath of hybridization.

Resumen

Las infecciones causadas por hongos son un problema sanitario creciente, y provocan ~1.350.000 muertes al año. Las especies del género *Candida* se encuentran entre los hongos patógenos más importantes. *Candida albicans* es la principal causante de infecciones por *Candida*, pero muchas otras especies del mismo género han emergido como patógenos. Los mecanismos evolutivos implicados en la adquisición de patogenicidad se desconocen, pero estudios precedentes apuntan a que la hibridación puede haber jugado un papel importante en este desarrollo. Los híbridos son organismos quiméricos que pueden presentar características fenotípicas únicas y una mayor plasticidad, lo que posiblemente facilite su adaptación a nuevos hábitats, como el cuerpo humano.

Esta tesis estudia las características genómicas de las especies patógenas del género *Candida*, centrándose principalmente en las especies híbridas y su evolución. Específicamente, se analiza la presencia de híbridos entre las especies de *Candida* y se estudian los procesos que impulsan la evolución de sus genomas. Para ello, se analizaron y compararon los genomas de 141 cepas correspondientes a 13 especies con el propósito de reconstruir sus características genómicas y estudiar su evolución. La naturaleza híbrida de seis de estas especies, incluida *C. albicans*, se describe por primera vez en esta tesis. Estos resultados sugieren que los híbridos están ampliamente distribuidos entre las especies de *Candida*. Todavía no conocemos qué factores podrían propiciar esta tendencia a la hibridación. Sin embargo, nuestros resultados indican que estos híbridos solo sufren incompatibilidades genómicas leves, y que la deriva genética es el factor dominante en su evolución. En resumen, esta tesis respalda un papel importante de la hibridación en la aparición de nuevas levaduras patógenas y aporta nuevas ideas sobre las consecuencias evolutivas de dicha hibridación.

Preface

In the last decades, advances in medical treatments and devices have increased our life-expectancy. However, through the use of invasive methodologies (e.g. catheters) or drugs leading to a weakening of the immune system (e.g. immunosuppressants), medical advancements have also increased our vulnerability to opportunistic infections. This notion is confirmed by the increasing incidence of yeast infections over the last twenty years, with *Candida* species having a relevant contribution, and representing an important burden in the hospital environment. As fungi were recognized only recently as an important threat for our health, the knowledge on their evolution, mechanisms of infection, or even response to medical treatments, is not as broad as for other pathogens. In this context, the European Training Network "OPATHY: From Omics to Patients" (www.opathy.eu) was created with the mission to exploit Next Generation Sequencing technologies for the study of *Candida* pathogens and their interaction with the human host, develop new diagnostics tools, and monitor the occurrence of yeast infections in the clinic. This thesis project is one of the thirteen projects of the OPATHY consortium and was responsible for the analysis of the genomes of important *Candida* pathogens, with a special focus on their evolution through hybridization. A brief overview of the different chapters of this thesis is presented below.

This thesis is divided in three main parts, namely: Introduction, Results, and Discussion. The first chapter of the Introduction, **Chapter 1**, comprises an overview of the epidemiology of *Candida* infections, the phylogenetic diversity of these pathogens, and genome variability. Although little is known about the evolution of these pathogens, hybridization has been suggested as an evolutionary mechanism for the emergence of new lineages with relevance to the clinic. These associations are explored in **Chapter 2**, which comprises a review article published during this project. In **Chapter 3**, a brief description of the different Bioinformatics techniques applied to Genomics are described with a particular focus on the analysis of hybrid genomes, which are the main target of this thesis. **Chapter 4** details the main objectives of the project.

In the frame of OPATHY network, the genomes of three *Candida* pathogens emerging in Europe were analyzed. The analyses of these genomes are reported in **Chapter 5**, with a genome report of two homozygous species, and **Chapter 6** with the extensive analysis of *Candida inconspicua* resulting in the finding that this pathogenic lineage originated via hybridization. To gain further insights on the evolution of hybrid *Candida* pathogens, we analyzed additional *Candida* hybrid and non-hybrid species. This analysis, which results are reported in **Chapter 7**, revealed the presence of more hybrid lineages, supporting the scenario that *Candida* species are prone to hybridize. The genome of *Candida albicans*, the major fungal pathogen, was also inspected to assess whether it presented characteristic features of hybridization. As reported in **Chapter 8**, the results suggest a hybrid ancestor at the origin of this species. To study genome evolution after hybridization and assess which factors could make *Candida* species so prone to hybridize, five natural hybrid lineages from *Candida parapsilosis* species complex were used as a model to compare hybrid genome evolution following different events of hybridization between the same parental species. This analysis is described in **Chapter 9**, and revealed that contrary to what was initially expected, these hybrids have weak genomic incompatibilities which may suggest that they are more prone to survive after hybridization. During this project, several gaps were found in the available tools for comparative genomics analysis, specifically with respect to the analysis of highly heterozygous genomes, as those of hybrids. In an attempt to contribute to fill in one of these methodological gaps, the HaploTypo pipeline was developed, and is described in **Chapter 10**.

As discussed in **Chapter 11**, this thesis project contributed to a better understanding of the evolution of *Candida* pathogens, and particularly of hybrid genomes. Furthermore, the large amount of data generated from this project represents an important resource for future studies.

Table of contents

Part I

Introduction

1 *Candida* yeast pathogens

1.1 *Candida* species and their pathogenicity

Candida species are a non-monophyletic group of budding yeasts that belongs to the Saccharomycotina subphylum (Figure 1_1) (Gabaldón et al., 2016). These species can be found in the environment, and some of them, such as *Candida albicans*, are also important commensals, being part of the normal human microbiota (Al-Yasiri et al., 2016; Bensasson et al., 2019; Consortium OPATHY & Gabaldón, 2019; Gabaldón & Fairhead, 2019; Ji et al., 2009; Papon et al., 2013; Sabino et al., 2015). Under certain circumstances, some of these species can adopt a pathogenic behavior that leads to an opportunistic infection (Consortium OPATHY & Gabaldón, 2019; Turner & Butler, 2014). Infections caused by *Candida* species, i.e. candidiasis, can vary from superficial infections, affecting cutaneous or mucosal epithelial cells (e.g. vaginal candidiasis), to severe life-threatening systemic infections (also known as candidemia) (Papon et al., 2013). It has been estimated that every year candidemia affects 250,000 people worldwide and causes 50,000 deaths (Kullberg & Arendrup, 2015). The most common agents of *Candida* infections are *C. albicans*, *Candida (Nakaseomyces) glabrata*, *Candida parapsilosis*, and *Candida tropicalis*, which together account for almost 90% of the cases of candidiasis (Pfaller et al., 2010b). It is important to note that the prevalence of each of these species depends on the geographical location, but *C. albicans* is, by far, responsible for the majority of the cases (>65%) (Pfaller et al., 2010b; Turner & Butler, 2014).

C. albicans is the major cause of vaginal fungal infections, but *Candida* species are also associated with hospital-acquired (nosocomial) candidiasis. Indeed, somewhat paradoxically, advances in medical treatments and devices have favored the occurrence of such infections to a point that *Candida* spp. are responsible for approximately 8% of the cases of hospital-acquired systemic infections (Consortium OPATHY & Gabaldón, 2019; Turner & Butler, 2014). A reason for this is the ability of some of these species to form biofilms (multicellular aggregates in which cells stick to each other and to surfaces) in medical devices like catheters, which are in

direct contact with the host (Finkel & Mitchell, 2011; Iraqui et al., 2004; Nobile & Mitchell, 2006). Furthermore, as *Candida* pathogens display an opportunistic behavior, the use of immunosuppressors or antibiotics, and any other factor that can contribute to a disequilibrium in the host microbiome or immune system, can also favor the occurrence of candidiasis (Consortium OPATHY & Gabaldón, 2019). For all these reasons, *Candida* species pose a particular risk for immunocompromised patients, such as transplant patients or those undergoing chemotherapy, neonates, and elderly people (Gabaldón & Carreté, 2016; Pfaller & Diekema, 2007).

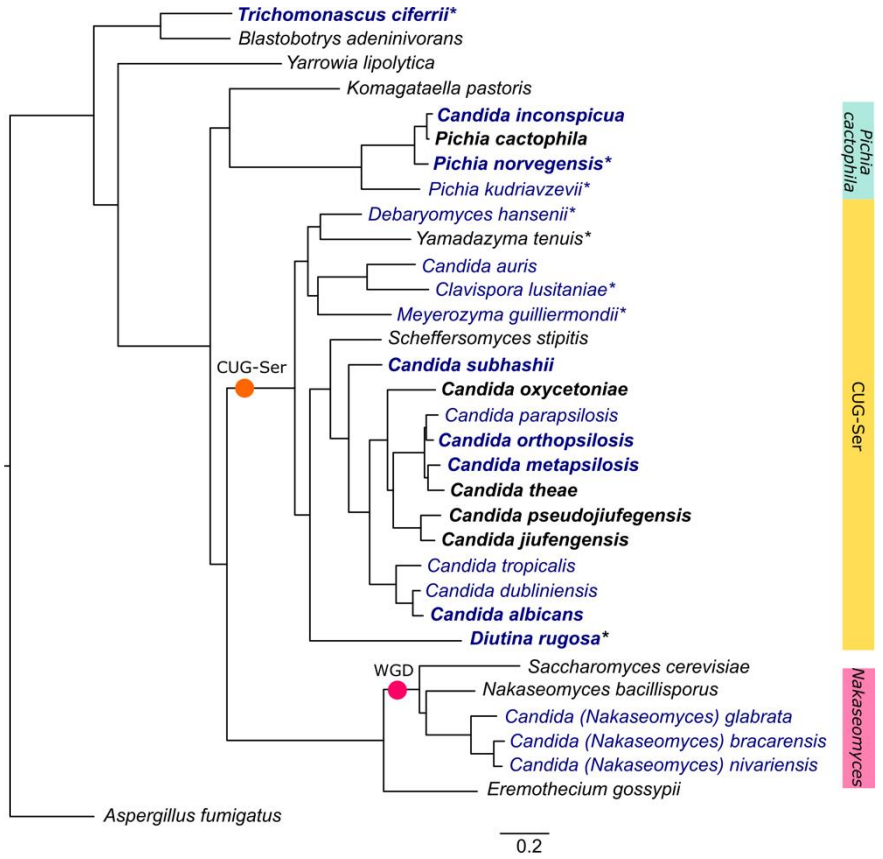

Figure 1_1. Phylogenetic tree with the main clades comprising *Candida* species. The three main clades with medical relevance are represented by green (*P. cactophila* clade), yellow (CUG-Ser clade) and pink (*Nakaseomyces* clade) bars. *Candida* pathogenic species are highlighted in blue, and those analyzed in this thesis in bold. An orange circle marks the CUG-Ser clade, while a pink circle marks a whole-genome duplication (WGD) event. Species marked with an asterisk were members of the *Candida* genus until very recently.

The classical methods for the diagnostics of *Candida* infections depend on the isolation and culture of the yeast species, making them time consuming (Consortium OPATHY & Gabaldón, 2019). This can represent an important drawback in the outcome of the disease. Therefore, efforts have been made for the development of new approaches which aim to exploit advances in sequencing and proteomics methodologies to improve the accuracy in the identification of the pathogen and reduce the time needed for a final diagnosis (Consortium OPATHY & Gabaldón, 2019). Such advancements will be essential to save patients' lives.

This need for new diagnostic methods is even more evident with the continuous broadening of the spectrum of possible agents of pathogenicity. Indeed, besides the outbreaks caused by the major *Candida* pathogenic yeasts (Lasheras et al., 2007; Nedret Koç et al., 2002; Pinhati et al., 2016; Qi et al., 2018; Thomaz et al., 2018), other *Candida* species are emerging as opportunistic pathogens, which is shifting the epidemiology of this disease (Blanco-Blanco et al., 2014; Pfaller et al., 2010b; Sabino et al., 2015; Trofa et al., 2008; Turner & Butler, 2014). For instance, *Candida auris* was able to achieve a worldwide distribution in less than 10 years, surpassing *C. glabrata* and *C. tropicalis* in the number of reported cases of candidemia (Sabino et al., 2020; Schelenz et al., 2016). Furthermore, other *Candida* yeasts, including "rare yeasts" (those with <0.1% prevalence) have been increasing their prevalence in the last years, and have also been responsible for outbreaks in hospitals, as is the case of *Diutina rugosa* (syn. of *Candida rugosa* (Khunnamwong et al., 2015)), or *Candida inconspicua* (D'Antonio et al., 1998; Guitard et al., 2013; Lopes Colombo et al., 2003; Pfaller et al., 2010b).

To further complicate this scenario, only four classes of antifungal drugs are currently available: azoles, echinocandins, polyenes and pyrimidine analogues. This not only limits therapy options, but also increases the chance of development of drug resistance (Ksiezopolska & Gabaldón, 2018). Indeed, there are reports of the development of resistance, or decreased susceptibility, to at least one class of antifungal drugs in some *Candida* species (Borman et al., 2019; Ksiezopolska & Gabaldón, 2018; Pfaller et al., 2010a). For example, *C. parapsilosis* azole-resistant strains were responsible for at least two recent outbreaks in Brazil (Qi et al., 2018; Thomaz et al., 2018), and the multi-drug resistant *C. auris* caused multiple fatal

outbreaks around the globe (Sabino et al., 2020). This scenario makes it imperative that studies embrace all the diversity of *Candida* pathogens, and not limit our knowledge to the model pathogen *C. albicans*. In this regard, it is important to understand not only their physiology and behavior as pathogens, but also their evolution and the mechanisms through which they emerged.

1.2 Phylogenetic diversity in *Candida* pathogens and mechanisms of pathogenicity

As mentioned before, *Candida* species belong to the Saccharomycotina subphylum. This subphylum is a subdivision of Ascomycota and is mostly known for harboring the model organism *Saccharomyces cerevisiae* (the baker's yeast) and *Candida* opportunistic pathogens. Members of Saccharomycotina can present different ecological behaviors, such as saprotrophism or parasitism. Moreover, they can be found almost everywhere around the globe, and in a broad variety of habitats, from deserts to fresh water (Kurtzman et al., 2011). Therefore, it constitutes a highly diverse group of species.

Although the term "*Candida*" is mostly associated to pathogenicity in the literature, this genus is a non-monophyletic group that also comprises multiple non-pathogenic species. Indeed, despite their shared genus name, *Candida* spp. comprise a group of highly diverged species, and not all of them are closely related (Figure 1_1). For instance, there are two main clades with clinical relevance, namely the CUG-Ser clade and the *Nakaseomyces* clade, which are genetically more distant than human and fish (Gabaldón & Carreté, 2016). These clades comprise both pathogenic and non-pathogenic species. The CUG-Ser clade includes the majority of *Candida* species, including *C. albicans*, *C. parapsilosis*, and *C. tropicalis*. Species of this clade have the particularity of translating the CUG codon as Serine, instead of Leucine (Santos et al., 2011). Additionally, the *Nakaseomyces* clade, which is closer to the baker's yeast *S. cerevisiae* than to the CUG-Ser clade, comprises *C. glabrata* and its relatives. *Nakaseomyces* species use the standard translating code and share with *S. cerevisiae* a past hybridization event (Gabaldón & Carreté, 2016; Marcet-Houben & Gabaldón, 2015).

Despite harboring the most prevalent *Candida* pathogens, together these two clades do not include all the species which are responsible for *Candida* infections. For example, *C. inconspicua*, which as mentioned before is an emerging pathogen, belongs to the *Pichia cactophila* clade, whereas *Trichomonascus ciferrii* (syn. *Candida ciferrii* (Kurtzman & Robnett, 2007)), which has recently been responsible for cases of infection (Gunsilius et al., 2001; Upadhyay et al., 2018), is a close relative of *Yarrowia lipolytica*, and consequently quite far from any of the above-mentioned *Candida* spp. (Figure 1_1).

The genomic variability present among *Candida* pathogens may have important repercussions in the development of their respective infections. Indeed, although the mechanisms of invasion of human cells are not broadly explored among *Candida* spp., they have been studied in *C. albicans* and *C. glabrata*, and it is known that they show many differences (Gabaldón & Carreté, 2016; Galocha et al., 2019). For instance, *C. albicans* infections start by the adhesion to the host epithelial cells through biofilm formation followed by invasion through an induced endocytosis, or, more commonly, through active physical penetration of cellular forms known as hyphae (Figure 1_2A) (Brunke & Hube, 2013; Galocha et al., 2019). Hyphae are able to pressure and disrupt the host cell membrane with the aid of hydrolytic enzymes, thus playing an important role in the invasion of epithelial cells and in the access to the bloodstream, favoring the occurrence of systemic infections (Brunke & Hube, 2013; Galocha et al., 2019; Goyer et al., 2016). Furthermore, they are also essential for *C. albicans* to escape from the host immune cells (Galocha et al., 2019). Regarding *C. glabrata*, the infection also starts with a process of adhesion to the host cells, but the invasion process is not fully understood. Contrary to *C. albicans*, *C. glabrata* does not form proper hyphal structures, and therefore the mechanism of invasion is possibly unrelated to physical pressure and associated to a process of endocytosis (Figure 1_2B) (Brunke & Hube, 2013; Galocha et al., 2019). However, more studies are needed to fully understand this mechanism.

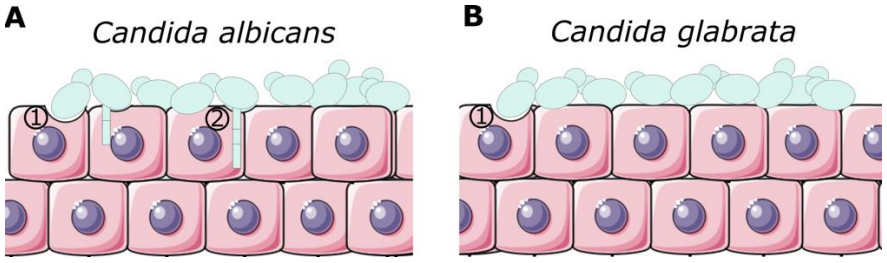

Figure 1_2. Schematic representation of the mechanism of invasion of human epithelial cells by **A)** *C. albicans* and **B)** *C. glabrata*. Number 1 shows the endocytosis of *Candida* yeast cells. This mechanism is present in *C. albicans* and possibly in *C. glabrata*, although the mechanisms of invasion of this last species are still not understood. Number 2 indicates active invasion by hyphae. This mechanism is not present in *C. glabrata.* Icons of epithelial cells were retrieved from smart.servier.com.

Although the specific mechanisms of invasion of host cells differ between *C. albicans* and *C. glabrata*, the overall pattern is similar: after detection of a possible host, cell adhesion is triggered and followed by invasion (Brunke & Hube, 2013). A previous study has proposed that the number of adhesin genes encoded in a genome might be related to the ability of a lineage to become a pathogen (Gabaldón et al., 2013). In *C. albicans* the essential adhesion proteins are encoded in Agglutinin-like sequence (ALS) adhesins and secreted aspartyl proteases (SAPs) genes, whereas in *C. glabrata* the adhesins are encoded in EPA genes (Brunke & Hube, 2013; Galocha et al., 2019; Hoyer et al., 2008). This genomic diversity underlying a similar trigger of pathogenicity is consistent with the previously proposed hypothesis that pathogenicity emerged multiple times independently (Gabaldón et al., 2013). It is important to note that adhesins are just a small piece of the puzzle. Other yet unknown factors can also be important. For instance, recently, a peptide toxin (candidalysin, encoded by the *ECE1* gene) secreted by hyphae was described as essential for invasion of human cells by *C. albicans* (Ho et al., 2020; Moyes et al., 2016; Naglik et al., 2019). The only two species with known *ECE1* orthologs are the closely related species *C. tropicalis* and *C. dubliniensis*. However, the levels of expression of *ECE1* are much lower in these two species than what is observed in *C. albicans* (Willems et al., 2018), suggesting that its role in virulence might be exclusive of this last species. These studies suggest, once again, that the large genomic variability among

Candida species has a great impact on their mechanisms of pathogenicity, complicating our understanding of these infections.

1.3 Mating and (a)sexual reproduction in *Candida* species

Species of the Saccharomycotina subphylum are budding yeasts, i.e. they have an asexual reproductive system through which a new cell grows at a specific site of the "mother" cell until they separate. This mechanism of clonal reproduction is widespread among species of this subphylum. Although asexual reproduction can be advantageous for the maintenance of favorable combinations of genes, the variability generated through sexual reproduction can present an advantageous fitness potential (Alby & Bennett, 2010). Therefore, some of the species of Saccharomycotina may also present a sexual cycle.

The best studied sexual reproductive system among these species is the one of the model organism *S. cerevisiae*. The most important loci for *S. cerevisiae* sexual reproduction are encoded in chromosome III and correspond to the silent *HMRa* and *HML*α loci, and to the active *MAT* locus (Figure 1_3A) (Haber, 2012). The mating type of a cell is defined by the idiomorph present in the *MAT* locus, which can either be **a** (comprising a single regulator gene) or $\boldsymbol{\alpha}$ (comprising two regulator genes, $\alpha 1$ and $\alpha 2$). In *S. cerevisiae*, mating occurs between two haploid cells with different mating types, which will produce a diploid cell able to sporulate. This is important because besides the genomic variability acquired through this process, spores are able to resist more extreme conditions (Huang & Hull, 2017). *S. cerevisiae* can experience mating-type switching, which consists of the exchange of the idiomorph present in the *MAT* locus by the alternative one (Butler et al., 2004; Haber, 2012). During switching the HO endonuclease cuts at the *MAT* locus, and this is repaired by recombination with *HMRa* or *HML*α, resulting in the replacement of the previously existing idiomorph. This mechanism ensures that even in a population where all cells have the same mating-type, there is the chance to undergo sexual reproduction.

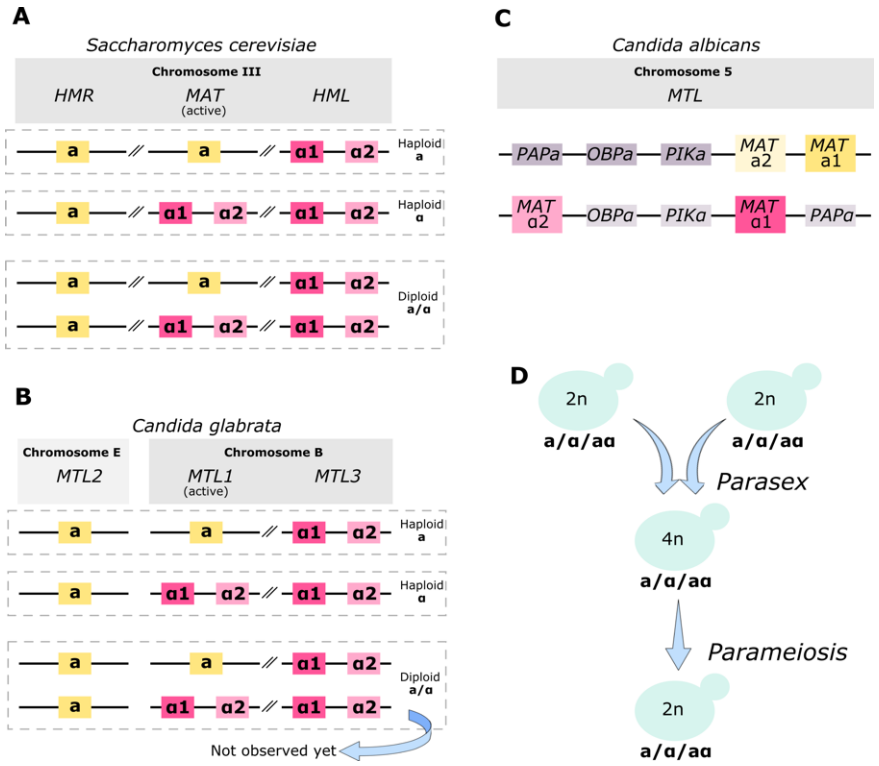

Figure 1_3. Schematic representation of the *MAT* locus of **A)** *S. cerevisiae*, **B)** *C. glabrata*, and **C)** *C. albicans.* Yellow and pink boxes represent the **a** and **α** idiomorphs, respectively. The additional genes of *C. albicans MAT* locus are in grey scale. **D)** Scheme of *C. albicans* parasexual cycle followed by parameiosis.

Until very recently, *C. glabrata* was thought to be an asexual organism. However, the analysis of its genome revealed the presence of homologs of all the genes involved in *S. cerevisiae* sexual reproduction (Fabre et al., 2005). Contrary to *S. cerevisiae*, the sexual reproduction machinery present in *C. glabrata* is divided in two chromosomes, namely chromosome B, which comprises *MTL1* and *MTL3* (homologs of *S. cerevisiae MAT* and *HMLα*, respectively), and chromosome E, which comprises *MTL2* (homolog of *S. cerevisiae HMRa*) (Figure 1_3B). However, so far none was able to induce mating in *C. glabrata* strains in the laboratory (Brisse et al., 2009; Gabaldón & Fairhead, 2019). Nevertheless, previous studies have provided evidence for the existence of both mating types and mating-type switching in *C. glabrata* populations, and, more recently, purifying selection was described to occur in these genes (Brisse et

al., 2009; Carreté et al., 2018; Gabaldón & Fairhead, 2019; Muller et al., 2008). All this suggests that *C. glabrata* is actively undergoing sexual reproduction (Gabaldón & Fairhead, 2019). Therefore, more studies are needed to assess the exact conditions under which *C. glabrata* is able to mate.

The absence of evidence for the occurrence of mating and meiosis in diploid *Candida* spp. of the CUG-Ser clade has led to the hypothesis that these species are asexual (Alby & Bennett, 2010). However, in the late 90s, a locus similar to *S. cerevisiae MAT* locus was described in *C. albicans* chromosome 5 (Hull & Johnson, 1999). As in *S. cerevisiae* and *C. glabrata*, the *MAT* locus of *C. albicans* also presents two idiomorphs, **a** and **α**, but there is an additional transcription regulator in the **a** idiomorph. Furthermore, besides the two regulators in each idiomorph, the locus is extended by three other genes, namely, *PIK*, *OBP* and *PAP* (Figure 1_3C), which play an important role in biofilm formation, but whose association to mating is not yet understood (Alby & Bennett, 2010; Hull & Johnson, 1999; Srikantha et al., 2012). Importantly, different experimental studies have shown that *C. albicans* cells are able to mate, and this can happen not only between cells with different idiomorphs, but also between cells with similar ones (Alby & Bennett, 2010; Hull et al., 2000; Magee & Magee, 2000). Contrary to *S. cerevisiae*, in *C. albicans* mating occurs by the cross of diploid cells. This mechanism is called parasexual cycle, because it differs from a normal sexual cycle not only in the ploidy of the mating cells, but also in the way the chromosome number is subsequently reduced (Figure 1_3D). After mating, a tetraploid cell is formed and the diploid state is not achieved through meiosis, but rather through a concerted chromosome loss (Bennett, 2015; Bennett & Johnson, 2003; Forche et al., 2008). This mechanism was recently described as parameiosis, because it recruits a machinery similar to that used in a normal meiotic process (Anderson et al., 2019). Parameiosis favors the acquisition of genomic variability due to high recombination rates and promotes the appearance of aneuploidies (the number of homologous chromosomes varies between chromosomes) (Anderson et al., 2019). Moreover, it avoids sporulation, which can be beneficial for a commensal or pathogenic species, because it circumvents a reaction of the host immune system to the presence of spores (Robinson, 2008). Although the remaining diploid *Candida* species of the CUG-Ser clade have a similar *MAT* locus to that of *C. albicans*,

besides this species, parasex has only been observed in *C. dubliniensis* and *C. tropicalis* (Butler et al., 2009; Porman et al., 2011; Pujol et al., 2004; Seervai et al., 2013). Further studies are needed to clarify how widespread parasex is among the CUG-Ser clade.

1.4 Genomic variability and recombination processes

The acquisition of genomic variability is a fundamental aspect of evolution. Genomic modifications (i.e. mutations) can contribute to the emergence of new phenotypes. Each mutation results in a different genotype that can be associated to a phenotype with a different fitness in a particular environment, and thus it is subject to the filter of selection. Mutations can appear through different mechanisms, for example, from the chemical instability of the nucleotide bases, or from copying errors during DNA replication. Mutations give rise to the presence of genetic polymorphisms within a population, species, or clade. These can consist of single nucleotide polymorphisms (SNPs), insertions or deletions (INDELs), copy number variations (CNVs), and genomic rearrangements. All these different types of polymorphisms can be detected through comparative genomics approaches and detecting them is essential to define differences between two lineages.

Genetic recombination is an important source of genomic variability. This process consists of the exchange of genomic material between different chromosomes, resulting in haplotypes that may differ from the original ones. Therefore, the outcome of such a process is more evident if there is a significant sequence divergence between the two recombining molecules. Genetic recombination is usually induced by DNA double strand breaks (DSBs), which are subsequently repaired using a homologous region as template (Dayani et al., 2011). As the donor region is typically residing in the paired homologous chromosome, this leads to the exchange of genetic information between them. In rare occasions recombination may occur between non-homologous chromosomes, resulting in translocations. DSBs repair occurs mostly through synthesis-depending strand annealing (SDSA) and double-Holliday junctions (DHJ) (Sung & Klein, 2006).

DHJ imply the formation of special DNA junctions in which four double strand arms are joined, while SDSA DNA repair occurs through annealing with the template sequence. For this reason, SDSA always results in interchanged regions that are flanked by the original genomic material (non-crossovers). In contrast, DHJ outcome depends on which mechanism resolves the junctions (Sung & Klein, 2006), possibly resulting in a non-crossover (similarly to SDSA) or a crossover event (the exchange of genetic material is extended until the end of the chromosome) (Dayani et al., 2011). Both cases result in a re-assortment of haplotypes which differ from the initial ones. A third possibility is nondisjunction, in which case homologs are not properly assorted into the daughter cells, resulting in the occurrence of aneuploidies, i.e. the number of homologs is not the same for all the chromosomes. It is important to note that recombination does not always imply reciprocal interchange of genomic material, as in some cases it may originate a non-reciprocal interchange, resulting in a gene conversion or loss of heterozygosity (LOH) event.

Recombination can be classified into mitotic or meiotic according to when it occurs within the cell cycle. Contrary to what the name suggests, mitotic recombination does not always occur during mitosis, as it can also occur during the interphase (LaFave & Sekelsky, 2009). In mitotic recombination, despite the possibility of occurrence of homologous chromosomal recombination, DSBs are normally solved with sister chromatids, i.e. copy of the exact same chromosome generated after DNA replication (Dayani et al., 2011; Symington et al., 2014). Therefore, the acquisition of genomic variability through inter-homologs mitotic recombination is very rare. This makes it difficult to induce and study inter-homologs mitotic recombination in laboratory (Dayani et al., 2011). In species with sexual reproduction, meiosis is the major source of chromosomal recombination. For a proper separation of chromosomes in the two daughter cells, it is necessary to ensure a proper homologous chromosome pairing. Therefore, during meiosis, DSBs are induced by the Spo11 enzyme and molecular junctions are formed between the homologous chromosomes, which are thereby linked together (Aguilera & Rothstein, 2007; Dayani et al., 2011; Petronczki et al., 2003). This promotes recombination and the interchange of homologous genomic material at a rate which is much higher than that of mitotic recombination (Brion et al., 2017).

As mentioned above, the outcome of recombination is more evident when it occurs between molecules with high sequence divergence. This is particularly relevant in highly heterozygous genomes as those of hybrids. Hybrids are chimeric organisms comprising the genome of two diverged lineages. Therefore, they present pairs of homeologous chromosomes, i.e. chromosomes that are derived from different parentals and are related by ancestry (Comai, 2005). Thus, contrary to the situation in organisms with low levels of heterozygosity, chromosomal recombination in hybrids is more likely to result in a large haplotype alteration with great impact on fitness. As hybrids may have an important role in the emergence of pathogens (Mixão & Gabaldón, 2018), this is further explored in the next chapter.

2 Hybridization and emergence of virulence in opportunistic human yeast pathogens

Mixão, V., & Gabaldón, T. (2018). Hybridization and emergence of virulence in opportunistic human yeast pathogens. *Yeast*, *35(1)*, 5-20. doi: 10.1002/yea.3242

2 Hybridization and emergence of virulence in opportunistic human yeast pathogens

2.1 Abstract

Hybridization between different species can result in the emergence of new lineages and adaptive phenotypes. Occasionally, hybridization in fungal organisms can drive the appearance of opportunistic lifestyles or shifts to new hosts, resulting in the emergence of novel pathogens. In recent years, an increasing number of studies have documented the existence of hybrids in diverse yeast clades, including some comprising human pathogens. Comparative and population genomics studies performed on these clades are enabling us to understand what roles hybridization may play in the evolution and emergence of a virulence potential towards humans. Here we survey recent genomic studies on several yeast pathogenic clades where hybrids have been identified and discuss the broader implications of hybridization in the evolution and emergence of pathogenic lineages.

Keywords: comparative genomics; emergence of virulence; hybridization; pathogens; yeast

2.2 Introduction

Inter-species hybrids result from the crossing of two diverged species. Hybrids are thus chimeric organisms that carry material from two differentiated genomes and may display a range of properties present in either of the two parent lineages, as well as novel, emerging phenotypes that differentiate them from both of their parents. As such, hybridization has been long recognized as an important evolutionary mechanism that can drive new adaptive phenotypes that enable the colonization of new environments (Gladieux et al., 2014). Studies on hybridization have been traditionally performed on

animals and plants, where hybrids are easy to recognize from their morphology. Such studies have established hybridization as an important evolutionary force driving the origin of new lineages and important ecological adaptations (Fonseca et al., 2004; Lee et al., 2013; Lunt et al., 2014; Masuelli et al., 2009; Session et al., 2016). In fungi, where hybrids are difficult to recognize morphologically or physiologically, the study of hybrids has long been neglected. However, the advent of genomics has recently enabled the identification of a growing number of fungal hybrids belonging to diverse clades and among industrial, clinical or environmental strains (Leducq et al., 2016; Morales & Dujon 2012; Pryszcz et al., 2014, 2015). Although most identified hybrids are thought to have been formed relatively recently, hybridization may have played an important role in the origin of some ancient lineages. Indeed, hybridization has recently been recognized as the process underlying the major whole-genome duplication that occurred around 100 million years ago in the lineage leading to *Saccharomyces cerevisiae* (Marcet-Houben & Gabaldón, 2015).

Besides its ecological and evolutionary importance, the study of hybridization is important from a physiological perspective. The situation created by the merging of two genomes, and the resulting transcriptomes and proteomes within a single cell, has been described as a 'genomic shock' (McClintock, 1984). In particular, at the early stages, physiological processes in hybrid organisms are expected to be intrinsically unfit, owing to interferences in the crosstalk of two genetic systems that have evolved separately for some time. Thus, even if some of the emerging new properties in the hybrid ensure a selective advantage in a given niche, the hybrid fitness will benefit from purging existing deleterious interactions between the two sub-genomes. This process is referred to as genome stabilization, and is mediated by several mechanisms including genomic recombination, gene conversion, or chromosomal loss or duplication (Payseur & Rieseberg, 2016). Although events mediated by such mechanisms have a stochastic mutational source, they can be subject to selection. Thus, evolution of fungal hybrid lineages follows particular rules, and we currently lack sufficient understanding of its tempo and mode.

In the last decade, the incidence of fungal infections has increased, partly owing to recent advances in the medical sector such as the increased survival of immunocompromised patients, the extensive

use of antibiotics, immunosuppressors or medical devices such as catheters (Gabaldón & Carreté, 2016; Pfaller & Diekema, 2007). This increase in incidence is due not only to a higher number of reported cases caused by known pathogenic species, but also an increasing number of infections whose underlying cause is a rare species (Pfaller & Diekema, 2004). For instance, among the etiological causes of invasive candidiasis, there are over 30 different species (Gabaldón et al., 2016), of which more than half can be considered very rare (i.e. <0.1% of the cases). This has led to the concept of emerging fungal pathogens (Papon et al., 2013). Although the above-mentioned progress in the medical sector is certainly one of the factors driving the increase in incidence and number of etiological agents, it is unclear whether other factors such as species dispersion to new environments may play a role. In addition, although the number of fungal species causing infection is increasing, there are clear differences among them in terms of their ability to cause infection and damage to the human host. In summary, we still have a very poor understanding of the genomic properties that may underlie a virulent outcome between a potential pathogen and its host, and how they have emerged during evolution.

In recent years hybridization has emerged as a potential important source of new pathogenic species. Indeed, a growing number of studies report the presence of hybrids among clinical isolates, in particular in yeast clades such as *Cryptococcus* or *Candida* (Boekhout et al., 2001; Pryszcz et al., 2014, 2015; Schröder et al., 2016; Viviani et al., 2006). These findings are worrisome, particularly considering current processes such as global trade, environment alteration and climate change, which may be favoring the encounter of divergent species that can still hybridize. The sources and potential implications of hybrids among clinical isolates are still poorly understood, and hence there is a need for studying the distribution, evolution and physiology of pathogenic hybrids. In this review we survey current knowledge about hybrids in opportunistic human pathogens, with a special emphasis on their genomic features and the mechanisms underlying their evolution. The possible relationship between hybridization events and the development of virulence will also be discussed.

2.3 Hybridization and the origin of emerging phenotypes

Hybrids are chimeric organisms carrying two different genomes that must coexist within the same cell. This phenomenon does not occur frequently in nature because on many occasions the chimeric organism is simply not viable. In the first instance mating between two organisms may be prevented by physiological prezygotic barriers such as problems in gamete recognition. If mating is possible and a zygote is formed, numerous post-zygotic barriers can exist. For instance, developmental problems may arise, aborting the generation of a new individual or, if the hybrid is viable, it may not be able to reproduce. Ultimately, the hybrid will need to survive in competition with other species. Certainly, all natural hybrid species that we now recognize are a small fraction of those that were initially formed, and they all must present some characteristics that promoted their survival.

Hybridization events can promote the origin of extreme phenotypes and adaptations to new ecological environments (Gladieux et al., 2014). This adaptation to new environments is key for hybrids' survival, as it not only allows isolation from the parent species but also it may provide a competitive advantage that allows for a relatively high fitness, despite the potential deleterious effects of the underlying 'genomic shock' (see below). Adaptation to new niches is generally promoted by the presence in the hybrid of transgressive phenotypes, that is, those phenotypes or characteristics that are beyond those present in the parent species. The sunflower *Helianthus paradoxus*, a hybrid between *Helianthus annuus* and *Helianthus petiolaris* is an example of this, as it can colonize soils with a high concentration of salt, where neither of the two parents can survive (Welch & Rieseberg, 2002a, 2002b). This possible existence of a superior performance of hybrids when compared with their parent lineages was described in 1908 by Shull, who later named this characteristic 'heterosis' (Shull, 1908; Shull, 1914). Heterosis can result from multiple mechanisms (Lippman & Zamir, 2007; Swanson-Wagner et al., 2006), and some studies point to the existence of a correlation between the parent expression levels and the hybrid performance (Frisch et al., 2010; Thiemann et al., 2010). This ability of hybrids to adapt to new conditions and show

transgressive phenotypes has been highly exploited in agriculture and in fact several crops that we include in our diet are hybrids (Bevan et al., 2017; Lippman & Zamir, 2007; Warschefsky et al., 2014). Besides agriculture, industries such as beer and winemaking are also taking advantage of the different characteristics of hybrid organisms (Pretorius & Bauer, 2002). *Saccharomyces pastorianus* is a hybrid yeast which keeps the strong fermenting ability of its parent *S. cerevisiae* as well as the ability to survive under low temperatures of its parent *Saccharomyces eubayanus* (Gibson & Liti, 2015). These two characteristics are essential for lager beer production, and therefore this organism has had an advantage compared with each of the parents, being widely used in this industry. Similarly, several hybrids between *S. cerevisiae* and other *Saccharomyces* species such as *Saccharomyces kudriavzevii* present the ability to grow under ethanol and temperature stress (Belloch et al., 2009). This characteristic has been extensively explored by wine makers, who prefer to produce white wine under low temperatures, minimizing the loss of aromatic compounds, which makes these hybrids an important economic asset in this business (reviewed in Marsit & Dequin, 2015). These are two of many possible examples of the relevance of hybridization for our daily life. In fact, these industries are increasingly recognizing the high potential of hybrids in adapting to new conditions and developing advantageous phenotypes. Currently, the artificial generation of new hybrids with useful characteristics in the laboratory constitutes a promising innovation strategy (Krogerus et al., 2017).

How can hybrids achieve these advantages? In crops, the mechanisms underlying heterosis have been extensively studied, and gene expression changes were reported in intraspecific hybrids of rice, maize and wheat (Guo et al., 2006; He et al., 2010; Stupar & Springer, 2006; Swanson-Wagner et al., 2006; Wang et al., 2006). These changes can be additive or non-additive when, respectively, the value of gene expression is the mean of both parent lineages, or the expression of one parent is decreased while the other is increased (reviewed in Chen, 2013). However, despite all of the studies performed to understand these mechanisms, the underlying molecular basis of heterosis is still poorly understood. More studies are definitely needed in order to understand this question better.

2.4 Genomic impacts of hybridization

As chimeric organisms, hybrids combine the genetic material of the two parent lineages. This implies a certain degree of genetic divergence between the homeologous chromosomes, that will initially equal the divergence between the two parent species at the time of hybridization. The resulting high levels of genetic heterozygosity may directly impact the functioning of the cell. Indeed, as suggested by the Bateson–Dobzhansky–Muller model (Bateson, 1909; Dobzhansky, 1934; Muller, 1942), different proteins for the same biological process can be produced at the same time, generating some incompatibilities that can influence the survival or fertility of the organism. Therefore, selection will favor any change that removes negative epistatic interactions in these heterozygous genomes, thereby increasing the chances of survival. In addition, even in the absence of selection, genomic changes involving differential loss of chromosomal regions or gene conversion will inevitably result in loss of heterozygosity (LOH) in particular regions. The combination of these selective and neutral processes will result in a progressive shaping of the heterozygous genome, which would gradually lose heterozygosity, while increasing its genomic stability. Several processes can contribute to this stabilization (Figure 2_1). For instance, the duplication of the entire set of chromosomes through whole-genome duplication would restore proper pairing between chromosomes, thus restoring the ability to go through meiosis (Marcet-Houben & Gabaldón, 2015; Wolfe, 2015). Alternatively, duplication or loss of individual chromosomes, leading to chromosomal aneuploidies, can also enhance genomic stability (Wertheimer et al., 2016). More specific mechanisms may contribute as well to this shaping process, namely gene loss, not only through literally deletion of the genomic region, but also through pseudogenization (Albalat & Cañestro, 2016), or through gene conversion, a process where a DNA sequence from one chromosome substitutes the one of its homeologue chromosome so that the two regions become identical (McGrath et al., 2014). Events of gene loss and gene conversion lead to LOH (Figure 2_1). This LOH is highly associated with the evolution of hybrids (Li et al., 2012; Louis et al., 2012; Stukenbrock et al., 2012), being an extremely important mechanism leading to the reduction of genomic incompatibilities present in such organisms.

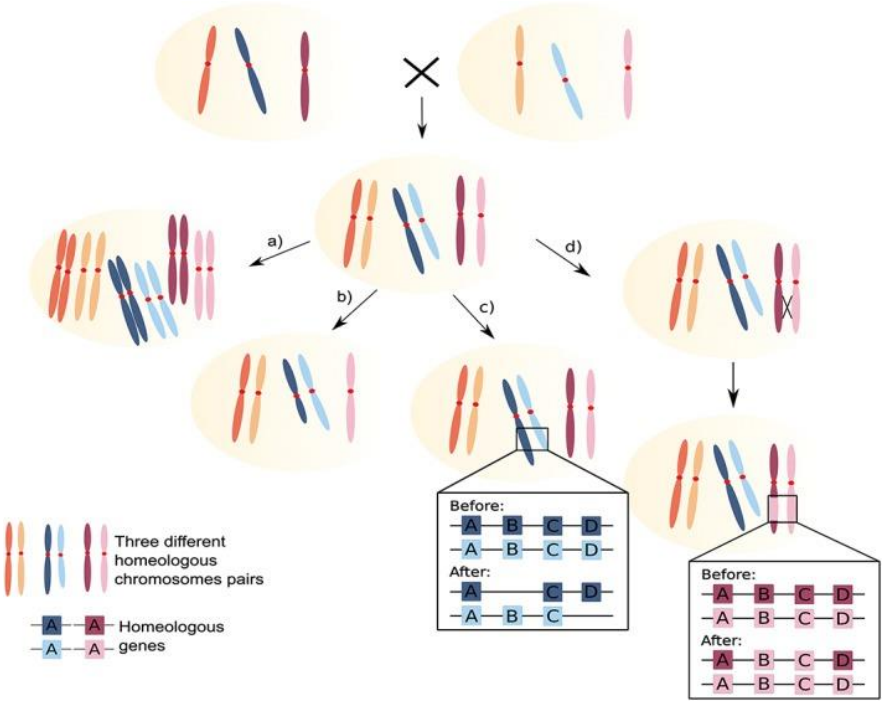

Figure 2_1. Schematic representation of several different mechanisms that can lead to genome stabilization in hybrids. Shaded ovals represent cells. Chromosomes are painted in different colors with different tones of the same color indicating homeologous pairs of chromosomes. In some cases, insets highlight specific regions with genes indicated as colored boxes. From top to bottom, two haploid cells from different species cross and form a diploid hybrid. Four non-exclusive, alternative evolutionary paths to genome stability are shown: (a) whole-genome duplication; (b) total or partial chromosome loss; (c) gene loss; and (d) gene conversion and loss of heterozygosity.

Another relevant factor for the stabilization of hybrid genomes is gene expression. Gene expression is regulated with *cis* (promoter region) and *trans* (elsewhere in the genome) elements, and in hybrid genomes it is possible that the *cis* component of one parent is regulated by the *trans* component of the other, which can alter gene regulation. As shown by Tirosh and colleagues, alterations in either *cis* or *trans* regulatory elements are important to obtain genomic stability, and therefore the reprogramming of gene expression should also be considered as an important factor in the stabilization of hybrid genomes (Tirosh et al., 2009). Furthermore, alterations in reprogramming of gene expression were shown to impact the way the

hybrid interprets sensory signals (Tirosh et al., 2009). This discovery raises the question whether all of these large alterations in hybrid genomes can have consequences in the way the organism interacts with the surrounding environment. It would be interesting in future works to address the question of whether these alterations in the signal perception can drive changes in the signal response, and consequently, for example, result in the virulence of a given hybrid pathogen.

2.5 Emerging pathogens and the evolutionary origins of virulence

Virulence results from the interaction of a microorganism with the host, as such it is an emerging property that depends on factors ranging from the genetic determinants of the host and the microorganism to the particular physiological or environmental conditions (Casadevall, 2012). According to Gabaldón and Carreté, not exclusively but particularly in opportunistic pathogens, virulence should be regarded as a secondary effect or an 'evolutionary accident', resulting from traits resulting from adaptations to selective pressure different from that involved in the pathogenic process itself (Gabaldón & Carreté, 2016). The number of fungal organisms with the ability to cause disease in humans may seem large, but it is only a small fraction of those that come into contact with us, as part of our microbiota or our environment. In addition, human fungal pathogens belong to evolutionary distinct clades and always have close relatives that are unable to infect humans (Figure 2_2). Thus, the ability to infect humans must have emerged several times independently. Uncovering what genomic changes may have promoted the emergence of virulence may serve to uncover novel virulence mechanisms. Despite recent efforts, we know very little about the evolution of virulence in fungi, although several trends are emerging, such as an increased cell-wall repertoire and adherence properties in pathogens as compared with closely related non-pathogens (Gabaldón et al., 2016). Hybridization provides an evolutionary scenario where radical genomic changes occur and where new emerging phenotypes may appear. If the new phenotype relates to the ability to survive in the human host, or an enhanced evasion from the immune system, we may obtain an increased virulence potential in a

possibly newly created lineage. The relationship between hybridization processes and the development of virulence potential is not well understood. However, hybridization is being recognized as an important path to virulence in the emergence of novel plant fungal pathogens (Depotter et al., 2016), and several human pathogenic species have been shown to present hybrids, such as *Candida orthopsilosis*, *Candida metapsilosis* or *Cryptococcus neoformans* (Boekhout et al., 2001; Pryszcz et al., 2014, 2015; Schröder et al., 2016).

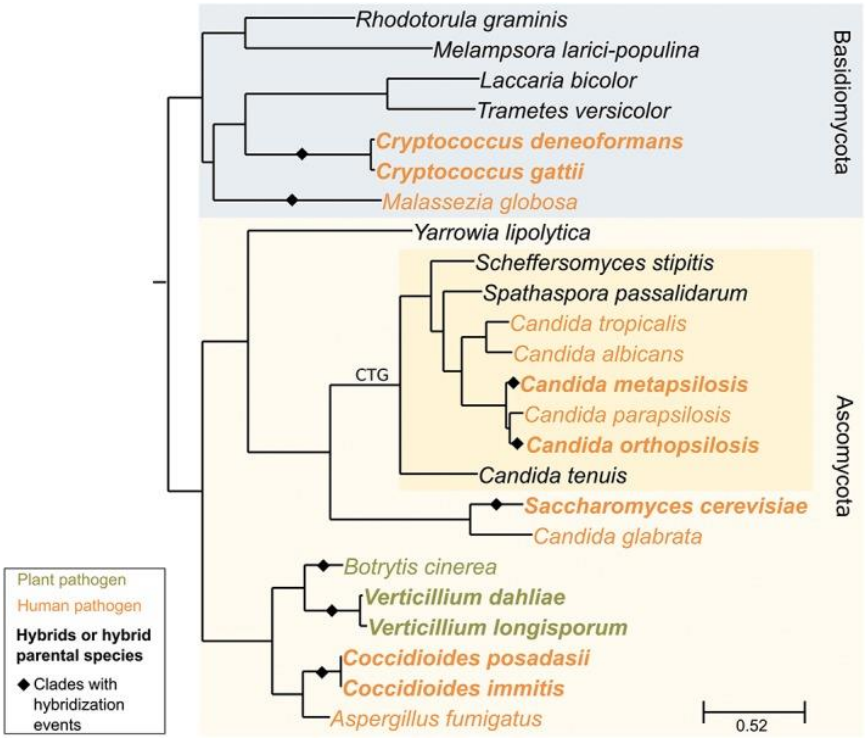

Figure 2_2. Evolutionary tree depicting the principal fungal clades with hybridization events in the origin of emergent pathogens. Clades where hybridization events have already been reported are indicated with the diamond symbol (◆). Hybrids or hybrid parent species are in bold. Basidiomycota and Ascomycota phyla are presented with blue and yellow backgrounds, respectively. Dark yellow highlights the CTG clade. Species already described as possible plant pathogens are in green, while those already described as possible human pathogens are in orange. The tree was reconstructed based on a set of four marker genes able to resolve fungal phylogenies (Capella-Gutierrez et al., 2014). Genes were aligned and trimmed following PhylomeDB pipeline (Huerta-Cepas et al., 2014), and the tree was reconstructed using the raxmlHPC-PTHREADS-SSE3 option of RAxML v8.2.4 (Stamatakis, 2014) set with PROTGAMMALG substitution model.

As discussed above, the initial instability present in hybrid genomes can represent a problem for these organisms as it has consequences in their fitness or survival. However, at the same time this instability results in high phenotypic plasticity and variability, which can represent an advantage in the adaptation to new environments. For example, the plant oomycete pathogen *Phytophthora* x*serendipita* is a hybrid originated through the mating of *Phytophthora cactorum* and *Phytophthora hedraiandra*, which infects the monocotyledon and dicotyledon species outside the host spectra of both parent species (Man in 't Veld et al., 2007). Thus, in the context of pathogenic organisms, hybridization can have an impact on the adaptation not only to different hosts, but also to different organs, as well as to different drugs.

The globalization and mobility of goods and people all over the world can contribute to the spread of new variants or species to new geographical locations. It is known that there is less reproductive isolation between organisms that are geographically distant, and that the emergence of hybrid pathogens is often associated with the introduction of new microbes in a given area (Depotter et al., 2016). In addition, global climate change and other alterations of the environment are changing the potential geographic distribution of species, favoring the colonization of new locations and opening the possibility that previously isolated species come into contact. Therefore, it can be speculated that these interchanges may contribute to the present and future increase of the emergence of hybrid species. Below, we survey recent studies on hybrids from the main clades of human pathogenic yeasts.

2.6 *Cryptococcus neoformans* and *Cryptococcus gattii*

The *Cryptococcus neoformans/Cryptococcus gattii* species complex comprises a very heterogeneous set of basidiomycete species, and its taxonomic organization has been reformulated several times during the last decade. Until the beginning of the present century, only one formal species was recognized – *C. neoformans*, with three recognized varieties, *C. neoformans* var. neoformans (corresponding to the serotype D), *C. neoformans* var. grubii (serotype A) and *C.*

neoformans var. gattii (serotypes B and C). However, in 2001, the analysis of molecular markers indicated the existence of two separated species: *C. neoformans* var. neoformans and var. grubii and *C. gattii* (Boekhout et al., 2001). More recently, a taxonomic reclassification based on phylogenetic analyses was proposed, and up to seven different species are now being considered. The two varieties of *C. neoformans* correspond to *C. neoformans* (previous var. grubii) and *Cryptococcus deneoformans* (previous var. neoformans), and *C. gattii* was divided into five different species, namely, *C. gattii*, *Cryptococcus bacillisporus*, *Cryptococcus deuterogattii*, *Cryptococcus tetragattii* and *Cryptococcus decagattii* (Hagen et al., 2015).

All of these species can cause human cryptococcosis (Brown et al., 2012; Hagen et al., 2015). Infections from these yeasts generally start by fungal spore cells reaching the human lungs through inhalation from the environment, since their main reservoir is trees (Cogliati et al., 2016). Once in the lungs the fungus can develop into a potentially fatal infection in immunocompromised patients, particularly AIDS patients (Viviani et al., 2006). In recent years however, some *Cryptococcus* outbreaks have affected presumed immunocompetent hosts in USA, Canada and Australia (Byrnes & Heitman, 2009; Byrnes & Marr, 2011; Galanis et al., 2010; Pappas, 2013). From the lungs, *Cryptococcus* infections can spread to other parts of the human body, causing additional clinical complications such as meningoencephalitis (Brown et al., 2012). The number of cryptococcosis in non-HIV patients is increasing (Bratton et al., 2012; Henao-Martínez & Beckham, 2015) and it is estimated that this disease affects 1 million people every year, killing around 625,000 (Park et al., 2009). However, given that not all countries have data available (Viviani et al., 2006), this number could be an underestimate. In France the presence of 0.3 infections per 100,000 persons/year was reported, with a fatality rate of 15%, while in Germany, between 2004 and 2013, 491 cases were reported (Bitar et al., 2014; Smith et al., 2015).

Molecular studies on the genetic background of *Cryptococcus* samples revealed the presence of several inter-species hybrids (Figure 2_3a), from which the AD hybrids (*C. neoformans* × *C. deneoformans* hybrids) are the most studied (see Table 2_1 for more information). The presence of different hybridization events could

eventually point to a high level of similarity between the different *Cryptococcus* genomes. However, these species are highly divergent at the nucleotide level, *C. neoformans* and *C. deneoformans* having 7% divergence, and *C. gattii* and *C. deneoformans* 13% divergence. Moreover, *C. deuterogatii* has 7.6% nucleotide divergence when compared with *C. gatti*, and 14.5% when compared with *C. deneoformans* (Table 2_1) (D'Souza et al., 2011; Janbon et al., 2014). The presence of hybrid species in *Cryptococcus* infections is a reality, with up to 30% of all infections being associated with AD hybrids (Viviani et al., 2006). Indeed, some hybrid strains were shown to present an increased virulence, suggesting that hybridization is associated with that trait (Hagen et al., 2015; Li et al., 2012). This represents a special concern, particularly if we take into account the fact that these pathogens are already revealing some heterosis as well as resistance to antifungal drugs (Li et al., 2012). Some studies point to the existence of more than one hybridization event in the origin of some *Cryptococcus* hybrids (Li et al., 2012; Xu et al., 2002), whose genomes are experiencing continuous chromosomal changes (Li et al., 2012; Rhodes et al., 2017), possibly reflecting the above-mentioned process of genome stabilization, but at the same time facilitating the path for pathogen evolution. Genomic analyses of AD hybrids revealed the presence of several aneuploidies, with some chromosomes from a specific parent being preferably retained (Hu et al., 2008). For instance, in chromosome 1, whose aneuploidies have already been associated with drug resistance (Sionov et al., 2010), there is a preference to retain the serotype A copy, which is sometimes duplicated, while its homeologue from serotype D is lost (Hu et al., 2008; Rhodes et al., 2017). It is notable that, for eight AD hybrid strains recently analyzed, a common origin was proposed at the same time that they were shown to have undergone different evolutionary paths (Rhodes et al., 2017). This is representative of the high plasticity of these genomes, and the high diversity generated after one hybridization event, contributing to the evolution of pathogenicity in these organisms.

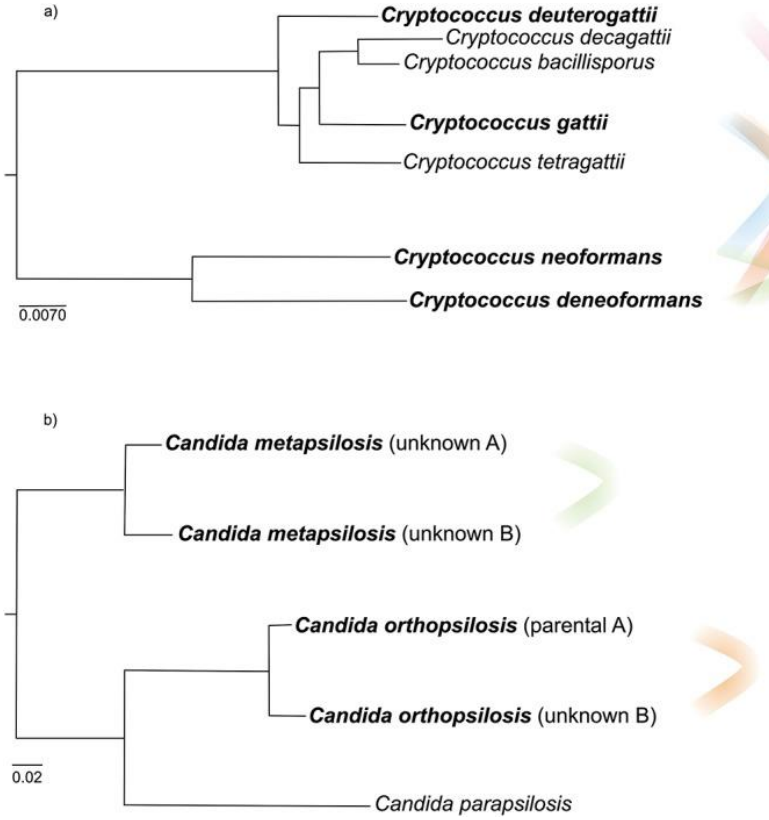

Figure 2_3. Schematic trees with representation of the different hybridization events already described in (a) *Cryptococcus* and (b) *Candida parasilosis* s.l. clades. For each tree, hybrid parent species are in bold and connected with colored curve lines. In each tree, different colors represent different hybridization events. The tree representing *Cryptococcus* clade was adapted from Hagen et al. (2015), while the one representing *C. parapsilosis* clade was adapted from Pryszcz et al. (2015).

Table 2_1. Recognized hybrid human pathogenic yeasts, and respective parent information, as well as evidence of the occurrence of hybridization events. Columns indicate, in this order: hybrid Phylum; hybrid name; first parent species; second parent species; genomic sequence divergence at the nucleotide level between the parent species; type of analysis used for hybrid detection; and literature where this information was retrieved from.

Phylum	Hybrid	Parental A	Parental B	Div	Evidence	Reference
Asco	*Candida orthopsilosis*	*Candida orthopsilosis*	Unknown	5%	Genome	(Pryszcz et al., 2014; Schroder et al., 2016)
Asco	*Candida metapsilosis*	Unknown	Unknown	4.5%	Genome	(Pryszcz et al., 2015)
Asco	*Coccidioides immitis* × *Coccidioides posadasii*	*Coccidioides immitis*	*Coccidioides posadasii*	n.a.	Genome	(Neafsey et al., 2010)
Asco	*Fusarium keratoplasticum*	*Fusarium keratoplasticum*	FSSC 9	n.a.	Markers	(Short et al., 2013, 2014)
Basidio	AD hybrids	*Cryptococcus neoformans*	*Cryptococcus deneoformans*	7%	Genotype	(Boekhout et al., 2001; Janbon et al., 2014; Rhodes et al., 2017)
Basidio	BD hybrids	*Cryptococcus deneoformans*	*Cryptococcus gattii*	13%	Genotype	(Bovers et al., 2008; D'Souza et al., 2011)
Basidio	AB hybrids	*Cryptococcus neoformans*	*Cryptococcus gattii*	n.a.	Genotype	(Bovers et al., 2008)
Basidio	*Cryptococcus neoformans* × *Cryptococcus deuterogattii*	*Cryptococcus neoformans*	*Cryptococcus deuterogattii*	n.a.	Genotype	(Aminnejad et al., 2012)
Basidio	*Malassezia furfur*	*Malassezia furfur*	*Malassezia furfur*	n.a.	Markers and genome	(Theelen et al., 2001; Wu et al., 2015)

Asco, Ascomycota; Basidio, Basidiomycota; Div, genomic divergence at nucleotide level; n.a, information not available; Genome, genomic analysis; Markers, genetic markers.

Some *Cryptococcus* pathogenic species appear as coexisting in the same niche (Cogliati et al., 2016), showing that they are constantly in contact and exposed to the possibility of occurrence of other hybridization events. The number of different hybrid species in this clade, and thus the apparent ease with which hybrids are formed recursively from independent events between the same lineages, is remarkable. This may be indicative of some special predisposition for these species to cross. This makes imperative the study of these hybrid genomes in order to understand their origin and genomic aftermath.

2.7 *Candida parapsilosis* complex

The *Candida parapsilosis* complex comprises three recently recognized species (Figure 2_3b) with a worldwide distribution: *Candida parapsilosis*, *C. orthopsilosis* and *C. metapsilosis* (Tavanti et al., 2005). All of these species are facultative commensals of human skin or mucosae, and all have the ability to form biofilms in catheters and other medical devices, being associated with opportunistic nosocomial infections (Lattif et al., 2010; Melo et al., 2011; Trofa et al., 2008). Hence, these species represent a special concern for AIDS, surgery and cancer patients, and other patients with long-term use of venous treatment. These opportunistic pathogens can be associated with episodes of fungemia, endocarditis, peritonitis, endophthalmitis, otomycosis and less frequently meningitis and vulvovaginal or urinary infections (Trofa et al., 2008). The treatment of these infections has to take into account their lower susceptibility to echinocandins as well as the possible presence of resistance to fluconazole (Garcia-Effron et al., 2008; Lockhart et al., 2008), which constitute the usual drugs used in candidemia treatment.

The human pathogenicity observed in this clade evolved independently from that in other *Candida* spp. (Pryszcz et al., 2015). Even so, *C. parapsilosis* ranks as the third most common cause of candidiasis worldwide, after *C. albicans* and *C. glabrata*, being the *Candida* species with the highest increase in incidence in recent years (Diekema et al., 2012; Trofa et al., 2008). *C. parapsilosis* is the most prevalent of the three species of the complex, and its infections particularly affect neonates, which account for one-third of *C.*

parapsilosis infections and have a relatively high mortality rate of around 10% (Chow et al., 2012; Pammi et al., 2013). Like *C. parapsilosis*, *C. orthopsilosis* can cause damage in human tissues, the less virulent member of this complex being *C. metapsilosis* (Gácser et al., 2007). These latter two species are less prevalent than *C. parapsilosis*, but their incidence varies greatly depending on the location. It is important to note that not all of the microbiology laboratories make the distinction between the three species of this complex (Trofa et al., 2008). A recent study performed between January 2009 and February 2010 in Spain reported that 29.1% of fungemia cases were caused by *C. parapsilosis* s.l.. From these, 8.2% were *C. orthopsilosis* and 1.1% *C. metapsilosis* (Cantón et al., 2011). However, in a survey performed in hospitals in China, this scenario was inverted with *C. metapsilosis* accounting for 11.3% of all *C. parapsilosis* s.l. infections, and *C. orthopsilosis* only 3.7% (Xiao et al., 2015). In fact, geographical location could be an important factor in the distribution and incidence of opportunistic organisms indirectly, because medical procedures and the clinical environment differ across the world.

The elucidation of the evolutionary history of the *C. parapsilosis* complex has uncovered the existence of several hybrids related to *C. orthopsilosis* and *C. metapsilosis* (Table 2_1). *C. parapsilosis* s.l. are diploid organisms and mating or meiosis has never been observed in this complex (Butler et al., 2009; Logue et al., 2005; Sai et al., 2011). *C. parapsilosis* is described as a highly homozygous species, as recent whole-genome analysis of various strains has confirmed (Butler et al., 2009; Pryszcz et al., 2013). In contrast, genomic analyses of *C. orthopsilosis* and *C. metapsilosis* provided a very different picture. Although the first genome sequence of a *C. orthopsilosis* strain depicted a highly homozygous sequence (Riccombeni et al., 2012), further genomic sequences revealed the existence of hybrids that had spread over distant geographical regions (Pryszcz et al., 2014; Schröder et al., 2016). The analysis of these hybrid genomes indicated a hybridization between the previously sequenced homozygous lineage and an unknown parent that showed 5% divergence at the nucleotide level (Pryszcz et al., 2014). Furthermore, it has been suggested that such hybrids have been formed at least four times independently, following independent hybridization events between the same parent lineages, and that the

majority of *C. orthopsilosis* strains are hybrids (Pryszcz et al., 2014; Schröder et al., 2016).

C. metapsilosis genome is highly heterozygous and, contrary to what is described for *C. orthopsilosis*, all of the sequenced strains of this species are hybrids that originated after a single hybridization event, and nothing is known about the parent species, except that they are 4.5% divergent (Pryszcz et al., 2015). The same authors hypothesize that, probably, the absence of the parent species in clinical samples is indicative of their inability to colonize or infect humans, and that the hybridization between these two non-pathogenic lineages originated an opportunistic pathogen with worldwide distribution.

Although not as high as in *Cryptococcus*, the presence of a high number of hybrids in this clade is notable. This leads to the question of whether, for some reason, this clade has a greater propensity to generate hybrids than other *Candida* species. More studies are needed to confirm this idea, but in contrast to the case of *Cryptococcus* hybrids, the genomic aftermath of hybridization has been extensively studied in hybrids of *C. metapsilosis* and *C. orthopsilosis* (Pryszcz et al., 2014, 2015; Schröder et al., 2016). An important observation regards the absence of differential loss of genomic material of the two parent species. In other studied hybrid species, the stabilization of the genome generally drives to a disequilibrium between the proportion of genetic material of the parent lineages in the hybrid. This phenomenon has been observed, for example, in *S. pastorianus* genome, where sometimes one of the parent genomes is kept while the other one is almost entirely lost (reviewed in Morales & Dujon, 2012). In the more recent hybrid *Meyerozyma sorbitophila*, where 40.3% of the genome has undergone LOH, a parent imbalance is already apparent with 88.8% of the LOH sequence corresponding to the preferred parent sub-genome (Louis et al., 2012). However, in the *C. parapsilosis* clade the proportion of the genetic material from each of the parent species in the LOH regions in the hybrids is close to 50% (Pryszcz et al., 2014, 2015), showing that possibly these genomes do not present strong deleterious incompatibilities as in other clades. As mentioned before, LOH is an important indicator of genome shaping and stabilization. The number of LOH events in *C. metapsilosis* was shown to differ between strains, with some events being shared between all of them (Pryszcz et al., 2015). Even so, *C. metapsilosis*

presents highly heterozygous regions in >50% of the genome, which is in contrast to the 17% of heterozygosity described for the MCO456 strain of *C. orthopsilosis* (Pryszcz et al., 2014, 2015). Owing to these differences, Pryszcz and colleagues proposed the existence of two possible scenarios: *C. metapsilosis* hybridization occurred more recently, or for some reason *C. orthopsilosis* genome is evolving faster (Pryszcz et al., 2015). The later work from Schroder and colleagues, showed that this MCO456 strain is associated with the oldest hybridization event of *C. orthopsilosis*, presenting one of the lowest heterozygosity values among the hybrid strains (Schröder et al., 2016). Indeed, within *C. orthopsilosis* it is possible to find strains with almost quadruple the heterozygous variants found in MCO456 (Schröder et al., 2016). Although these studies have been performed, several questions are still pending. For instance, it is as yet unknown whether the LOH events occur preferably at genomic regions encoding certain genes or functions, which could imply that these genes or functions are essential for the hybrid's survival. Answering this question would open the door to a better understanding not only of the genome stabilization phenomenon, but also of the special ability for the species of this complex to hybridize and the emergence of their ability to infect humans.

2.8 Hybrids in other human fungal pathogens

The relevance of hybridization to human health is not exclusive to *Candida* or *Cryptococcus* clades. Gene exchange through hybridization was already reported in other species. Coccidioidomycosis is a pulmonary infection caused by the inhalation of *Coccidioides immitis* or *Coccidioides posadasii* spores, which are present in the soil and air. Although not being a dangerous disease in immunocompetent patients, it is estimated that 150,000 new cases of coccidioidomycosis occur annually in the USA, of which one-third are fatal, being a special concern for immunocompromised persons (Odio et al., 2017). Genomic analyses of some populations from both species revealed the presence of a recent hybridization event (Neafsey et al., 2010), and more recently, the analysis of some genetic markers uncovered again the presence of hybrid organisms in these populations (Johnson et al., 2015). For instance, the genomic patterns of introgression between these two

species revealed that at least 8% of the genes in *C. immitis* population may have been recently introgressed from *C. posadasii*, presenting an enrichment in genes associated with immune evasion and cell walls (Neafsey et al., 2010). Based on these results, the authors have proposed that these antigenic genes may present a selective advantage when introduced in the genome of the other species, which raises concern about the consequences of this hybridization for the virulence of these pathogenic fungi.

The same concern is appearing in relation to the filamentous fungi *Fusarium keratoplasticum*, the main causative agent of nosocomial infections associated with plumbing systems. *Fusarium* is responsible for skin lesions, pneumonia and disseminated infections in immunocompromised patients (Nucci & Anaissie, 2007), and strains of these fungi have been reported to present resistance to some antifungal drugs (Azor et al., 2007; O'Donnell et al., 2008). Recent studies based on several genetic markers uncovered the presence of hybrid strains within this species, suggesting that this natural hybridization resulted in adaptation to the anthropogenic environments (Short et al., 2013, 2014). Similar reports have been done for *Malassezia furfur*, a commensal fungus of the human skin, sometimes responsible for skin disorders, such as eczema or atopic dermatitis (Theelen et al., 2001). Recent advances in genomics and bioinformatics enabled the confirmation of a hybrid origin in some of its strains, where genes, usually present in one copy in this species, are present in two copies (Wu et al., 2015). A further phylogenetic analysis allowed confirmation that the hybrid strains were originated by hybridization events between two highly distant lineages of *M. furfur* (Wu et al., 2015). Therefore, the authors suggest that they should be regarded as intra-species hybrids. Alternatively given the high distance between *M. furfur* lineages, if we consider *M. furfur* as a species complex, they could be considered as inter-species hybrids (Wu et al., 2015).

Altogether, these findings show that hybridization is a transversal phenomenon in fungal pathogens, occurring in different clades (Figure 2_2). Moreover, the advent of genomics is allowing the discovery of such organisms, as well as study of their evolution and possible characteristics. Hence, it is possibly fair to assume that the number of known hybrid pathogens is far from the real total, and fungal hybridization is more frequent than previously thought.

However, it is also important to note that the discovery of hybrids is just the first step towards understanding the evolution of hybrids in a clade and its possible role in the emergence of virulence. The examples described in this review are possibly only the tip of the iceberg. Of the five reported clades with hybrids, only two have been investigated to a certain degree, and constitute the most thoroughly studied cases. Yet studies on *Candida* and *Cryptococcus* hybrids have only superficially approached the mechanisms underlying hybrid genome evolution, and how they relate to phenotypic change, in particular with respect to their pathogenic behavior.

2.9 Concluding remarks and future prospects

Hybridization is a biological process responsible for the origin of new lineages or species with adaptation to new environments. The increasing identification of hybrid species with medical or economic importance shows that this evolutionary process has to be regarded as an important driver of evolution, especially in the fungal clade. Advances in medical procedures allowing the survival of immunocompromised patients, globalization and climate change are all factors probably underlying the increase in the number of hybrid opportunistic fungal pathogens. The trend is also likely to increase even further. Considering this, there is a need for investment in the development of new diagnostic tests in order to apply the optimal therapeutics. For this reason, it is important to understand the mechanisms underlying the hybridization events in fungi, especially those leading to the harmonious coexistence of two different genomes in one unique organism.

In this regard, the fields of genomics and bioinformatics are set to play a crucial role in future studies. In order to achieve this, the future will need to witness further developments of new programs and tools especially directed to the assembly and analysis of fungal genomes, with a particular focus on highly heterozygous genomes. This should be regarded as an urgent task since good genomic sequences are the essential basis of achieving better knowledge on these pathogens. Short-read paired-end technologies are widely used for genome sequencing. However, difficulties in solving complex genome assemblies are leading to an integrated use of multiple technologies

(Goodwin et al., 2016). Long-read sequencing produces reads with several kilobases, which can be useful to solve regions where ambiguities are present (Goodwin et al., 2016), thus representing an important advance for hybrid genome assemblies. Even so, the higher error rates associated with long-read sequencing are still a big concern. Although some works indicate that these are random errors, and therefore high coverage could overcome the problem, the fact is that these technologies can present an error rate up to 14% (Carneiro et al., 2012; Koren et al., 2012). Nevertheless, it is expected that new technological developments will progressively reduce error rates. Facing the problem of highly heterozygous assemblies, new pipelines such as 'Redundans' have recently been developed, contributing to an improvement in the quality of heterozygous genome assemblies (Pryszcz & Gabaldón, 2016). Nevertheless, there is still much room for improvement. For instance, 'Redundans' and other assembly programs reconstruct chimeric reference genomes, comprising interspersed regions of the two sub-genomes. Although subsequent mapping of genomic reads against this reference can distinguish homozygous from heterozygous blocks, in the absence of known parent species with a sequenced genome, it is not possible to reconstruct the origin of each of the sub-genomes. There exist solutions for genome phasing, but these were specially developed for bacteria or mammals, and not for yeasts. This gap may be compromising the study of hybrid pathogens, and consequently, for example, the evaluation of the possibility that hybridization events are responsible for the emergence of new pathogens. Phasing hybrid genomes would help understand this problem, since in some cases, as mentioned before for *C. metapsilosis*, both parent lineages are unknown (Pryszcz et al., 2015), making it difficult to assess the consequences of such hybridization. Therefore, for a better understanding of these hybrid pathogens, new bioinformatics strategies have to be developed, which certainly will contribute to an improvement of our knowledge on hybridization in fungi, and consequently to the clarification of the consequences of this phenomenon for human health.

Genomic analyses can certainly increase our understanding of the evolution of hybrid pathogens, yet they need to be complemented with experimental functional analysis. Only a few studies have compared pathogenesis-relevant phenotypes such as drug resistance, adherence or virulence among different hybrid strains and their

homozygous parents. Only the comprehensive and carefully planned phenotyping of a large number of hybrid and non-hybrid strains from which the genomes have been sequenced will help us understand the genetic mechanisms underlying virulent traits in hybrids. Finally, transcriptome analysis performed with technologies such as RNAseq has the potential to reveal how transgressive phenotypes may be achieved in yeast hybrids and how hybridization can rewire transcriptional networks.

In conclusion, hybridization is a process suggested to be at the origin of new emerging pathogens (Hagen et al., 2015; Li et al., 2012; Pryszcz et al., 2015). There is still much to learn about these organisms, as current studies are only helping us to realize the potential role of hybridization in the emergence of pathogenesis, but we are as yet far from understanding this process. Although the real dimensions of this phenomenon are currently unknown, the currently described cases are likely to be only the tip of the iceberg. In addition, the emergence of hybrids with virulence potential may be increasing thanks to global trade and climate change. Finally, infections from hybrids may be more difficult to treat as they may be able to adapt faster owing to their intrinsically high genomic plasticity. Thus, pathogenic hybrids can represent a special concern for human health, and their study should be a matter of interest to us.

3 Comparative genomics analysis of hybrid genomes

3.1 From sequencing reads to the analysis of hybrid genomes

Advances in Next Generation Sequencing (NGS) have facilitated the acquisition of whole-genome sequencing data. The increased accessibility that research laboratories have to this technology, particularly due to decreased costs, has shifted the way in which biological research is performed (Consortium OPATHY & Gabaldón, 2019). As mentioned in the previous chapter, nowadays, there are several sequencing strategies available, from which the two most used ones produce either high throughput and accurate short paired end reads, or less accurate long reads (up to 1Mb) at lower throughput. In all of them the final output consists of multiple sequences corresponding to fragments of genomic DNA ("reads"), together with information on the quality for the inference of each base in the sequence. These reads need to be filtered according to their quality to avoid errors introduced by the sequencing process. As reads correspond to small genomic fragments, it is necessary to infer their coordinates with respect to the whole-genome sequence. To this end, standard pipelines align those reads to an available genome sequence ("reference genome"), a process also known as "read mapping" (Figure 3_1). This approach provides a map with the coordinates of each of the genomic reads of the analyzed sample, and the differences with respect to the reference sequence, which can then be used to assess the genomic variability by a comparative genomics analysis (details in next section).

As mentioned in the previous chapters, hybrid genomes have high levels of heterozygosity. This can have important repercussions in downstream bioinformatics analyses due to the presence of two possible haplotypes (in a diploid hybrid) for each heterozygous genomic region (Figure 3_1A). Therefore, it is important to identify those genomes early in the analysis. The K-mer Analysis Toolkit (KAT, (Mapleson et al., 2017)) computes the density of the number of reads supporting the different *k*-mers in the sequencing data. As in

homozygous genomes the coverage is expected to be similar across the genome, such analysis usually outputs a normal distribution. Nevertheless, as heterozygous regions have two different *k*-mers (in a diploid), a bimodal distribution in which one peak has half of the coverage of the other is the expected result for highly heterozygous genomes. Hence, KAT (Mapleson et al., 2017)) is a useful tool to detect possible hybrids before further bioinformatics analyses.

Knowing the level of sequence divergence between the two haplotypes in hybrid genomes is crucial to decide the best analytical approach. For instance, if the divergence between the two haplotypes allows reads of one haplotype to align to the other one, the reference genome can be a so-called "reduced genome" in which only one of the haplotypes is represented for each homeologous region (Figure 3_1B) (Pryszcz & Gabaldón, 2016). In an alternative approach, which is the one to apply when the sequence divergence does not allow reads from one haplotype to align to the other, the reference genome should comprise the two haplotypes ("phased reference genome", Figure 3_1C). As the hybrid's parental genomes are not always known, sppIDer was recently developed with the aim to identify the possible parentals represented in each genomic region (Langdon et al., 2018). This tool requires multiple good genome assemblies for closely related species, where the hybrid sequencing reads are aligned.

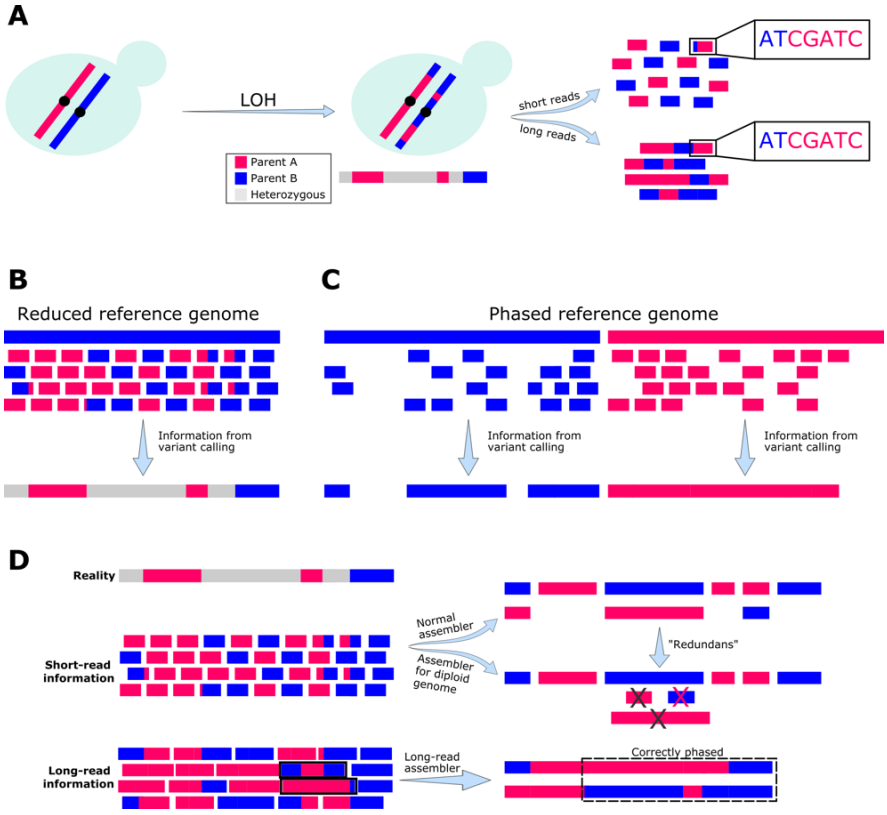

Figure 3_1. Implications of a hybrid genome for bioinformatics analysis. **A)** A hybrid yeast cell with two homeologous chromosomes (blue and red) undergoes loss of heterozygosity (LOH) events. The overall variability is represented by a bar under the evolved genome, in which regions with LOH maintaining the copy of parental A (red), parental B (blue) or both (grey, heterozygous) are represented. Genomic reads can be obtained by short- and long-read sequencing. Reads covering the extremities of LOH regions may have information of different haplotypes. **B)** Read mapping on a "reduced" reference genome. Reads of both haplotypes are aligned in the same region, and information on heterozygous and LOH regions is obtained. Nevertheless, in heterozygous regions, it is not known to which haplotype an allele belongs. **C)** Read mapping on a "phased" reference genome. Reads align in the most similar haplotype, allowing the assessment of genomic variability in each of them, including heterozygous regions. However, information on heterozygosity and LOH is not obtained. **D)** Representation of the real genomic features in terms of heterozygosity (grey) and LOH (red and blue), and indication of the short and long reads covering such regions. Long-reads spanning a LOH block are highlighted with a black box. Possible strategies of *de novo* genome assembly are represented by arrows. In regions where a long read spans a LOH block, the genome assembly is expected to be correctly phased.

The higher accessibility to sequencing technologies is promoting the sequencing of new species for which a reference genome is not available. In such cases, the genomic reads have to be assembled to derive a reference genome for the first time ("*de novo* genome assembly"). Hybrid genomes pose specific additional challenges to this process, especially if only short reads are available. The high levels of heterozygosity are expected to result in the two haplotypes assembling into different sequences ("contigs"). However, the presence of LOH regions which are longer than any of the available read pairs makes it difficult to connect the flanking haplotypes, resulting in a highly fragmented genome (Figure 3_1D). Different algorithms specially developed to deal with heterozygous genomes are available. This is the case the genome assembler dipSPAdes, or, as mentioned in the previous chapter, the post-assembly processing tool 'Redundans' (Pryszcz & Gabaldón, 2016; Safonova et al., 2015). 'Redundans' is a program designed to remove redundant contigs from the assembly and run additional steps to link the remaining contigs when possible. Therefore, this tool reduces the fragmentation of the genome assembly (Pryszcz & Gabaldón, 2016). In the end, a reduced genome assembly is obtained in which the short reads may be aligned. It is important to note that each contig in the assembly may correspond to a mix of haplotypes, and therefore it is difficult to know whether two given genes correspond to the same parental sub-genome. A combined approach of short- and long-read sequencing technologies may help to reduce the fragmentation of the assembly and is expected to decrease the mixing of different haplotypes (Figure 3_1).

3.2 Assessment of genomic variability

Genomic variability can be posed in terms of SNPs, deletions, insertions, or any other differences that a sequenced individual or population might have when compared to a reference genome. The analysis of all these features in a hybrid genome depends on the type of reference genome that was used, i.e. a reduced or a phased reference genome. After read mapping on a reduced reference genome, several public tools can be used to detect SNPs and INDELs (Hwang et al., 2015). These tools report for each variable position the genotype that was observed. In the case of heterozygous positions

(two alleles are present) the number of reads supporting each allele can be used to calculate the allele frequency and estimate the ploidy level of the sample. Furthermore, the density of heterozygous SNPs across the genome can also be used to determine heterozygous and LOH blocks, as well as to estimate the current haplotype divergence (Pryszcz et al., 2015). However, even if the information on all the heterozygous positions is retrieved, it is not possible to connect all heterozygous alleles belonging to a single haplotype, as the reference itself is an unresolved mixture of the two haplotypes. Programs like HapCUT2 (Edge et al., 2017) or WhatsHap (Martin et al., 2016) are available to phase (i.e. separate in current haplotypes) these variants. Nevertheless, the presence of long LOH blocks that exceed the size of a short read may once again represent a problem. Besides all the mentioned constraints, the reconstruction of ancestral haplotypes is further complicated by the presence of events of reciprocal chromosomal recombination.

In the case of phased reference genomes, where the two parental genomes are known, the genomic variability in terms of heterozygosity is difficult to assess, because usually there is not a one-to-one coordinate correspondence between the two homeologous chromosomes. For the same reason, the identification of regions of LOH is also more challenging (Figure 3_1). So far, there are no available tools able to perform any of these tasks. The advantage of using such a reference is the ability to assess phasing information, which allows the reconstruction of the current haplotypes of each hybrid strain, through the analysis of homozygous SNPs. This information is useful in a comparative genomics analysis between multiple closely related hybrid samples, because it allows the identification of the ancestral variability in each sample, which can provide a better picture of the underlying hybridization event(s).

In summary, the complexity of hybrid genomes still poses many challenges for Bioinformatics, which range from the genome assembly to the assessment of genomic variability. The above described approaches have advantages and drawbacks, with a reduced assembly making it more difficult the assessment of phasing information, and a phased reference complicating the identification of heterozygous and LOH regions. As a note, in this thesis project, a new pipeline was developed to get, when possible, the best of the two approaches (HaploTypo, Chapter 10).

4 Objectives

This thesis aimed to assess how widespread hybridization is among *Candida* spp., and to study the processes underlying the evolution of these hybrid genomes. To this end, specific objectives were defined. These objectives are listed below:

- Assess the frequency, distribution, and nature of hybridization events among *Candida* species using *de novo* genome assembly and comparative genomic approaches on emerging yeast pathogens and their close relatives.

- Evaluate the possibility that the heterozygosity patterns of *C. albicans* represent the footprints of an ancient hybridization event.

- Study the genomic aftermath of hybridization using as a model the genome sequences of natural hybrids resulting from independent crosses of the same two lineages.

- Analyze the genome of the first environmental isolate of the pathogenic species *C. metapsilosis.*

- Contribute with new tools for the analysis of heterozygous genomes.

Part II

Results

5 Genome Assemblies of Two Rare Opportunistic Yeast Pathogens: *Diutina rugosa* (syn. *Candida rugosa*) and *Trichomonascus ciferrii* (syn. *Candida ciferrii*)

Mixão, V., Saus, E., Hansen, A. P., Lass-Flörl, C., & Gabaldón, T. (2019). Genome Assemblies of Two Rare Opportunistic Yeast Pathogens: *Diutina rugosa* (syn. *Candida rugosa*) and *Trichomonascus ciferrii* (syn. *Candida ciferrii*). *G3: Genes|Genomes|Genetics*, *9*(12), 3921–3927. doi: 10.1534/g3.119.400762

5 Genome Assemblies of Two Rare Opportunistic Yeast Pathogens: *Diutina rugosa* (syn. *Candida rugosa*) and *Trichomonascus ciferrii* (syn. *Candida ciferrii*)

5.1 Abstract

Infections caused by opportunistic yeast pathogens have increased over the last years. These infections can be originated by a large number of diverse yeast species of varying incidence, and with distinct clinically relevant phenotypic traits, such as different susceptibility profiles to antifungal drugs, which challenge diagnosis and treatment. *Diutina rugosa* (syn. *Candida rugosa*) and *Trichomonascus ciferrii* (syn. *Candida ciferrii*) are two opportunistic rare yeast pathogens, which low incidence (< 1%) limits available clinical experience. Furthermore, these yeasts have elevated Minimum Inhibitory Concentration (MIC) levels to at least one class of antifungal agents. This makes it more difficult to manage their infections, and thus they are associated with high rates of mortality and clinical failure. With the aim of improving our knowledge on these opportunistic pathogens, we assembled and annotated their genomes. A phylogenomics approach revealed that genes specifically duplicated in each of the two species are often involved in transmembrane transport activities. These genomes and the reconstructed complete catalog of gene phylogenies and homology relationships constitute useful resources for future studies on these pathogens.

Keywords: *Candida ciferrii*; *Candida rugosa*; *Diutina rugosa*; *Trichomonascus ciferrii*; genome assembly; pathogen; yeast

5.2 Introduction

Candida species are the most common cause of hospital-acquired fungal infections, very often leading to patient's death (Brown et al., 2012; Gabaldón & Carreté, 2016; Lass-Flörl, 2009; Pfaller & Diekema, 2007). Although *Candida albicans*, *Candida glabrata* and *Candida parapsilosis* are the species with highest prevalence (Jordà-Marcos et al., 2007; Pfaller & Diekema, 2007), in the last years the incidence of "rare yeast" infections has increased (Bretagne et al., 2017; Lass-Flörl, 2009; Pfaller et al., 2012). By "rare yeasts" we mean ascomycetous yeasts that have very low prevalence (< 1% of clinical *Candida* infections) and have high Minimum Inhibitory Concentrations (MICs) toward at least one class of antifungal drugs (Bretagne et al., 2017; Guitard et al., 2013; Jung et al., 2015).

Diutina rugosa (syn. *Candida rugosa* (Khunnamwong et al., 2015)) and *Trichomonascus ciferrii* (syn. *Candida ciferrii* (Kurtzman & Robnett, 2007)) are two "rare yeasts" (Pfaller et al., 2010b). *D. rugosa* has been reported as a causative agent of veterinary infections (Crawshaw et al., 2005; Moretti et al., 2000; Scaccabarozzi et al., 2011), and therefore might have impact in industry and economics. Furthermore, it has been identified as the etiological agent of several clinical infections, including a clinical outbreak in Brazil (Lopes Colombo et al., 2003; Pfaller et al., 2010b). Thus, this species is considered an emerging fungal pathogen (Minces et al., 2009; Pfaller et al., 2006). Indeed, in a 10-year multi-center study a 10-fold increase in the number of *D. rugosa* clinical cases was reported (Pfaller et al., 2010b). *T. ciferrii* has also been reported as an opportunistic pathogen in some sporadic cases of infections in immunocompromised patients (Gunsilius et al., 2001; Pfaller et al., 2010b; Saha et al., 2013; Upadhyay et al., 2018; Villanueva-Lozano et al., 2016). Both species were recently shown to present high MICs to azoles and echinocandins (Pérez-Hansen et al., 2019; Pfaller et al., 2006).

Next Generation Sequencing (NGS) is a powerful tool to study the genomic background of pathogens, which might reveal many of their features. In the last years, more and more studies performing NGS analysis on yeast pathogens were published and showed the relevance of whole-genome sequence for the study of pathogenic genomic

determinants (Butler et al., 2009; Mixão et al., 2019; Pryszcz et al., 2015; Ropars et al., 2018; Schröder et al., 2016). In this context, we decided to sequence the genome of both *D. rugosa* and *T. ciferrii*, which will be useful for future studies on these opportunistic pathogens.

5.3 Materials and Methods

Library preparation and genome sequencing

We sequenced the type strains for *D. rugosa* (CBS613) and *T. ciferrii* (CBS4856). Genomic DNA extraction was performed using the MasterPure Yeast DNA Purification Kit (Epicentre, United States) following manufacturer's instructions and all reagents mentioned are from the kit if not specified otherwise. Briefly, cultures were grown in an orbital shaker overnight (200 rpm, 30°C) in 15 ml of YPD medium (Yeast extract-Peptone-Dextrose medium: 10 g of yeast extract, 20 g of bacto peptone and 50 ml of dextrose 40% in 1 L of distilled water). Cells were harvested using 4.5 ml of each culture by centrifugation at maximum speed for 2 min, and then they were lysed at 65°C for 15 min with 300 µl of yeast cell lysis solution (containing 1 µl of RNAse A). After being on ice for 5 min, 150 µl of MPC protein precipitation reagent were added into the samples, and they were centrifuged at 16.000 g for 10 min to pellet the cellular debris. The supernatant was transferred to a new tube, DNA was precipitated using 100% cold ethanol and centrifuging the samples at 16.000 g, 30 min, 4°C. The pellet was washed twice with 70% cold ethanol and, once the pellet was dried, the sample was resuspended in 100 µl of TE. All gDNA samples were cleaned to remove the remaining RNA using the Genomic DNA Clean & Concentrator kit (Epicentre) according to manufacturer's instructions. Total DNA integrity and quantity of the samples were assessed by means of agarose gel, NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, United States) and Qubit dsDNA BR assay kit (Thermo Fisher Scientific).

Whole-genome sequencing was performed at the Genomics Unit from Centre for Genomic Regulation (CRG) with an Illumina HiSeq2500 machine. Libraries were prepared using the NEBNext

Ultra DNA Library Prep kit for Illumina (New England BioLabs, United States) according to manufacturer's instructions. All reagents subsequently mentioned are from the NEBNext Ultra DNA Library Prep kit for Illumina if not specified otherwise. 1 μg of gDNA was fragmented by nebulization using the Covaris S2 instrument (Covaris Inc.) to a size of ~600 bp. After shearing, the ends of the DNA fragments were blunted with the End Prep Enzyme Mix, and then NEBNext Adaptors for Illumina were ligated using the Blunt/TA Ligase Master Mix. The adaptor-ligated DNA was cleaned-up using the MinElute PCR Purification kit (Qiagen, Germany) and a further size selection step was performed using an agarose gel. Size-selected DNA was then purified using the QIAgen Gel Extraction Kit with MinElute columns (Qiagen) and library amplification was performed by PCR with the NEBNext Q5 Hot Start 2X PCR Master Mix and index primers (12–15 cycles). A further purification step was done using AMPure XP Beads (Agentcourt, United States). Final libraries were analyzed using Agilent DNA 1000 chip (Agilent) to estimate the quantity and check size distribution, and they were then quantified by qPCR using the KAPA Library Quantification Kit (KapaBiosystems, United States) prior to amplification with Illumina's cBot. Libraries were loaded and sequenced in paired-end reads of 125bp on Illumina's HiSeq2500. Base calling was performed using Illumina pipeline software. In multiplexed libraries, we used 6 bp internal indexes (5' indexed sequences). De-convolution was performed using the CASAVA software (Illumina, United States). Sequence data has been deposited in short read archive (SRA) under the BioProject Accession No. PRJNA531406.

De novo genome assembly and phylome reconstruction

Raw sequencing data were inspected with FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).
Paired-end reads were filtered for quality below 10 or size below 31 bp and for the presence of adapters with Trimmomatic v0.36 (Bolger et al., 2014). The K-mer Analysis Toolkit v2.4.1 (KAT; (Mapleson et al., 2017)) was used to get the GC content and *k*-mer frequency distribution and estimate the expected genome size. SOAPdenovo v2.04 (Luo et al., 2012) was used to perform genome assembly. Redundant contigs were removed with Redundans v0.13c (Pryszcz & Gabaldón, 2016) using default parameters, i.e., 51% minimum

identity and at least 80% overlap. The quality of the assembly was inspected with Quast v4.5 (Gurevich et al., 2013) and KAT v2.4.1 (Mapleson et al., 2017). Species identification was confirmed by BLASTn (Zhang et al., 2000) of the respective ITS region (accession: NR_111249.1 for *D. rugosa* and NR_111160.1 for *T. ciferrii*), as recommended (Stavrou et al., 2018). Genome annotation was performed with Augustus v3.1 (Stanke & Morgenstern, 2005), using *Meyerozyma guilliermondii* and *Sacharomyces cerevisiae* as model organisms for *D. rugosa* and *T. ciferrii*, respectively. The Ascomycota dataset in BUSCO v3 (Waterhouse et al., 2018) was used to assess completeness.

Phylome reconstruction - i.e., the complete collection of phylogenies for every gene encoded in the genome - was performed using the PhylomeDB pipeline (Huerta-Cepas et al., 2014), as described in (Pryszcz et al., 2015), considering twenty-seven species (Supplementary file 1). This was done for both *D. rugosa* and *T. ciferrii*, using their respective predicted proteomes as seed. These phylomes and the corresponding orthology and paralogy relationships are available for browsing or download in PhylomeDB (Huerta-Cepas et al., 2014) with ID 932 and 842, respectively. Gene gain and loss analysis in seed branch was performed based on the phylome results. A BLASTp (Zhang et al., 2000) was performed against the UniProt database (UniProt Consortium, 2019) (accessed on April 30th, 2019), in order to determine the possible function associated with these genes, as well as their GO terms. An enrichment analysis was done using FatiGO (Al-Shahrour et al., 2007). Species-tree reconstruction was based on the final concatenated alignment of 469 single genes, comprising 297,788 amino-acid positions, with RAxML v8.2.4 (Stamatakis, 2014), using the PROTGAMMALG substitution model and performing rapid bootstrapping with 1,000 replicates. A BLASTp (Zhang et al., 2000) of the species-specific genes against the non-redundant database maintained by NCBI (NCBI Resource Coordinators, 2015) (accessed on September 13th, 2019), considering only hits with e-value < 0.001 and query coverage > 50%, was performed to determine whether these genes have homologs in species which were not considered for phylome reconstruction.

Read mapping and variant calling

Read mapping for all strains (Table 5_1) was performed with BWA-MEM v0.7.15 (Li, 2013). Picard v2.1.1 (http://broadinstitute.github.io/picard/) was used to sort the resulting file by coordinate, as well as to mark duplicates, create the index file, and obtain mapping statistics. Mapping results were inspected with IGV version 2.0.30 (Thorvaldsdóttir et al., 2013). Mapping coverage was determined with SAMtools v0.1.18 (Li et al., 2009).

Samtools v0.1.18 (Li et al., 2009) and Picard v2.1.1 (http://broadinstitute.github.io/picard/) were used, respectively, to index the reference and create a dictionary to be used in subsequent variant calling steps. GATK v3.6 (McKenna et al., 2010) was used to call and filter variants with the tools HaplotypeCaller and VariantFiltration, respectively, as described by (Mixão et al., 2019). In order to determine the number of SNPs/kb, a file containing only SNPs was generated with the SelectVariants tool. Moreover, for this calculation only positions in the reference with 20 or more reads were considered for the genome size, and these were determined with bedtools genomecov v2.25.0 (Quinlan & Hall, 2010).

Mitochondrial genome assembly

NOVOPlasty v2.7.2 (Dierckxsens et al., 2017) with default parameters was used to assemble *D. rugosa* and *T. ciferrii* mitochondrial genomes, taking as seed input the respective *Cox2* gene (accession numbers: KT832772.1 and DQ443088.1, respectively). The final assemblies were complete, as the assembly program was able to circularize each of them. Mitochondrial genome annotation was performed with MITOS2 (Bernt et al., 2013). Read mapping to these mitochondrial assemblies was performed as mentioned before for the nuclear genome.

Data availability

Data generated by this project can be found under the BioProject PRJNA531406, including sequencing data, genome assemblies and respective annotation. Phylomes can be found in PhylomeDB, with

the phylome IDs 842 and 932. A list of species used for phylome reconstruction, and the results of the enrichment analysis can be found in Supplementary files 1 and 2, respectively. Plots related to the *k*-mer analysis in the genome assemblies are in Supplementary figure 1. Supplemental material available at figshare: https://doi.org/10.25387/g3.8945048.

5.4 Results and Discussion

Genome sequencing and assembly

In this study we sequenced the type strains of *D. rugosa* and *T. ciferrii*, using an Illumina-based, pair-end sequencing strategy (see Materials and Methods). GC content and 27-mer count analyses of the sequencing reads revealed only one peak for each strain (Supplementary figure 1), suggesting that the two sequenced strains are highly homozygous. Based on the same 27-mer counts, we estimated genome sizes of approximately 13 Mb and 19 Mb for *D. rugosa* and *T. ciferrii*, respectively (see Materials and Methods). We next performed a *de novo* genome assembly for each of these species (see Materials and Methods). The final nuclear genome assembly of *D. rugosa* comprised 13.4 Mb, with 49.56% GC content and a N50 of 193,138 bp (Table 5_1). This assembly was divided in 171 contigs, of which 88 were longer than 25 kb, representing 97.7% of the genome. Automated gene prediction resulted in 5,821 protein-coding genes (see Materials and Methods). Despite the fact that the genome size was close to our estimations and similar to the one reported for the closely related species *Diutina catenulata* (13.1 Mb), the number of predicted proteins was substantially lower than the 7,128 proteins annotated in the close relative *D. catenulata* (O'Brien et al., 2018). Furthermore, only 64.07% of the reads could be mapped to *D. rugosa* nuclear genome assembly. These observations made us question the completeness of our assembly. However, KAT (Mapleson et al., 2017) reported that 98.96% of 27-mers was represented in the assembly (Supplementary figure 1), and BUSCO (Waterhouse et al., 2018) reported 97.7% completeness of *D. rugosa* predicted proteome. Finally, most of the reads that did not map to the nuclear genome were found to correspond to the mitochondrial genome (see section "Mitochondrial genome assembly"). Thus, we consider that

D. rugosa genome annotation is not significantly underestimating its gene content. It remains to be investigated whether the large number of proteins reported for *D. catenulata* is an annotation artifact or a real biological difference. There are no other available genomes for this genus, and the close relatives *M. guilliermondii* and *Scheffersomyces stipitis* have 5,920 and 5,841 annotated proteins, respectively (Butler et al., 2009; Jeffries et al., 2007).

Table 5_1. Metrics of *D. rugosa* and *T. ciferrii* nuclear genome assemblies, with indication of their respective genome size, N50, GC content, coverage, percentage of mapped reads, variants per kilo-base (kb) and heterozygous (heter) variants per kb.

Species (strain)	Size (Mb)	N50	GC (%)	Coverage (reads/ position)	Mapped reads (%)	SNPs / kb	Heter SNPs/ kb
Diutina rugosa (CBS613)	13.4	193 138	49.56	175.9	64.07%	0.09	0.07
Trichomonascus ciferrii (CBS4856)	20.5	69 012	47.46	209.6	98.35%	0.12	0.09

The nuclear genome assembly of *T. ciferrii* comprised 20.5 Mb, with 47.46% GC content and a N50 of 60,012 bp (Table 5_1). This assembly entailed 584 contigs, of which 132 were longer than 25 kb, representing 84.3% of the genome. Genome annotation predicted 6,913 proteins (see Materials and Methods). To the best of our knowledge, there is no other genome assembly of the *Trichomonascus* genus published so far, which would allow us to have a better assessment of the quality of our assembly. Even so, 27-mer frequency analysis showed that 99.83% of *T. ciferrii* 27-mers was represented in the assembly (Supplementary figure 1), and BUSCO (Waterhouse et al., 2018) estimated 93.4% proteome completeness, suggesting a good representation of the *T. ciferrii* genome in our assembly. Read mapping and variant calling confirmed that both *D. rugosa* and *T. ciferrii* are highly homozygous, having 0.07 and 0.09 heterozygous SNPs/kb, respectively (Table 5_1). It is worth mentioning that both genomes present homozygous SNPs (0.02 SNPs/kb in *D. rugosa* and 0.03 SNPs/kb in *T. ciferrii*), which is unexpected as the reads were mapped on the respective

assembly. This situation can probably be a result of errors introduced during the sequencing process or data analysis (i.e., read assembly, read mapping or variant calling).

Mitochondrial genome assembly

As mentioned before, only 64.07% of *D. rugosa* sequencing reads mapped to the respective genome assembly. Thus, we decided to assemble its mitochondrial genome (see Materials and Methods), in order to see whether the remaining reads could come from it. A final 41.8 kb circular mitochondrial genome assembly was obtained, suggesting that the assembly is complete (Figure 5_1A). Read mapping confirmed that 34.5% of *D. rugosa* sequencing reads corresponded to the mitochondrial genome, which suggests a high mitochondrial content in this yeast. We have assembled *T. ciferrii* mitochondrial genome as well, obtaining a circular assembly with 29.2 kb (Figure 5_1B), where 2.2% of *T. ciferrii* sequencing reads mapped. While we annotated 14 protein-coding genes in *D. rugosa* mitochondria, in *T. ciferrii* we annotated 16. The major difference between the two species involved the *nad4L* (associated to complex I) and *rps3* genes, which were absent in *D. rugosa*.

Comparative genomics

In order to elucidate particular characteristics of *D. rugosa* and *T. ciferrii* we decided to follow a comparative genomics approach and compared their nuclear genomes/proteomes with other species. We reconstructed the complete collection of gene evolutionary histories (i.e., the phylome) (Gabaldón, 2008) for each of these two species, in the context of twenty-six other species (see Material and Methods, Supplementary file 1). We identified 770 species-specific genes for *D. rugosa* and 1,217 for *T. ciferrii*, from which only 247 and 391, respectively, had homologs in species which were not considered for phylome reconstruction (see Materials and Methods). In both species, species-specific genes were not enriched in any particular function. Interestingly, genes specifically duplicated in each of the two species seemed to be enriched in transmembrane transport activities, as well as, oxidoreductase activity (detailed information can be found in Supplementary file 2). As can be observed in the species tree (Figure

5_2A), *D. rugosa* belongs to the CUG-Ser1 clade, while *T. ciferrii* is close to *Yarrowia lipolytica*. This shows that although very distantly related these two emergent pathogens present gene duplications affecting similar functions. Furthermore, in the case of *D. rugosa*, it is worth noting an enrichment in aspartic-type endopeptidase activity and ferrichrome transporter activity (Supplementary file 2), as both have been reported as important for pathogenic behavior, particularly in *Candida* species (Heymann et al., 2002; Naglik et al., 2003).

An earlier study on *D. catenulata* genome revealed an interesting break in the *MAT* locus of this species (O'Brien et al., 2018). By comparison with the *MAT* locus of *M. guilliermondii* (Reedy et al., 2009), these authors reported the absence of *PAP* gene close to *MAT alpha1* (Figure 5_2B), being the *PAP* gene instead in a different contig of *D. catenulata* genome assembly (O'Brien et al., 2018). Furthermore, they found that this gene was phylogenetically closer to *PAP a* than to *PAP alpha* (O'Brien et al., 2018). To assess whether this characteristic is shared within the *Diutina* genus, we here inspected the *MAT* locus of *D. rugosa*. Contrary to *D. catenulata* (O'Brien et al., 2018), *D. rugosa MAT* locus corresponded to the *MAT a* allele, where, similarly to *M. guilliermondii*, we could only find *MAT a2* (Figure 5_2B). Moreover, when comparing to *M. guilliermondii* (Reedy et al., 2009), we could not identify any particular rearrangement in this locus. Regarding the *MAT* locus of *T. ciferrii*, we observed that, in contrast to *Y. lipolytica* where both *MAT a* and *MAT alpha* were described (Butler et al., 2004), it only presents the *MAT alpha* allele (Figure 5_2b). It is worth to mention that although there is a protein-coding gene in the place of *MAT alpha2*, this protein does not present any homolog and therefore we were only able to identify *MAT alpha1* in *T. ciferrii* genome (Figure 5_2b).

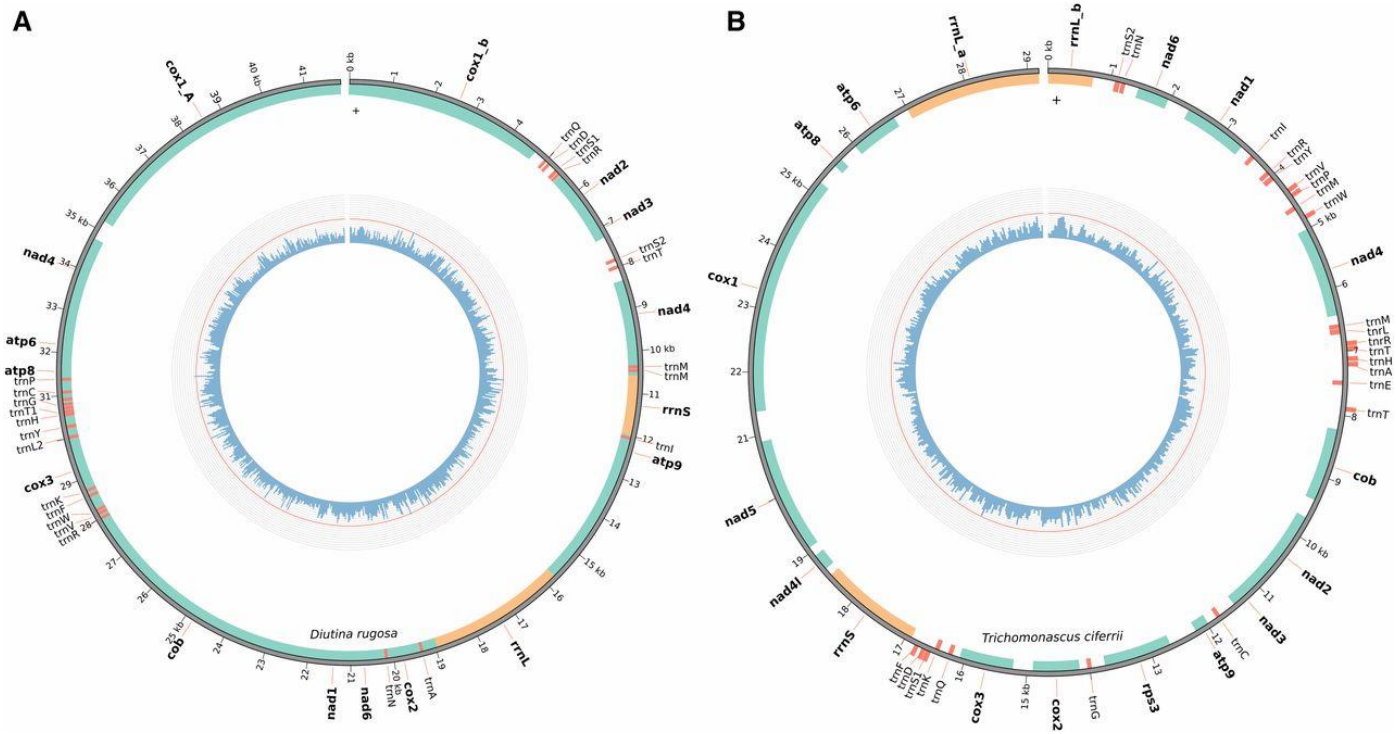

Figure 5_1. Mitochondrial genome representation of **A)** *D. rugosa* and **B)** *T. ciferrii*. Protein-coding genes are marked in green, tRNA genes are marked in red, and ribosomal genes are marked in orange. The blue histogram in the center represents the GC content variation. A more legible version of this figure is available online.

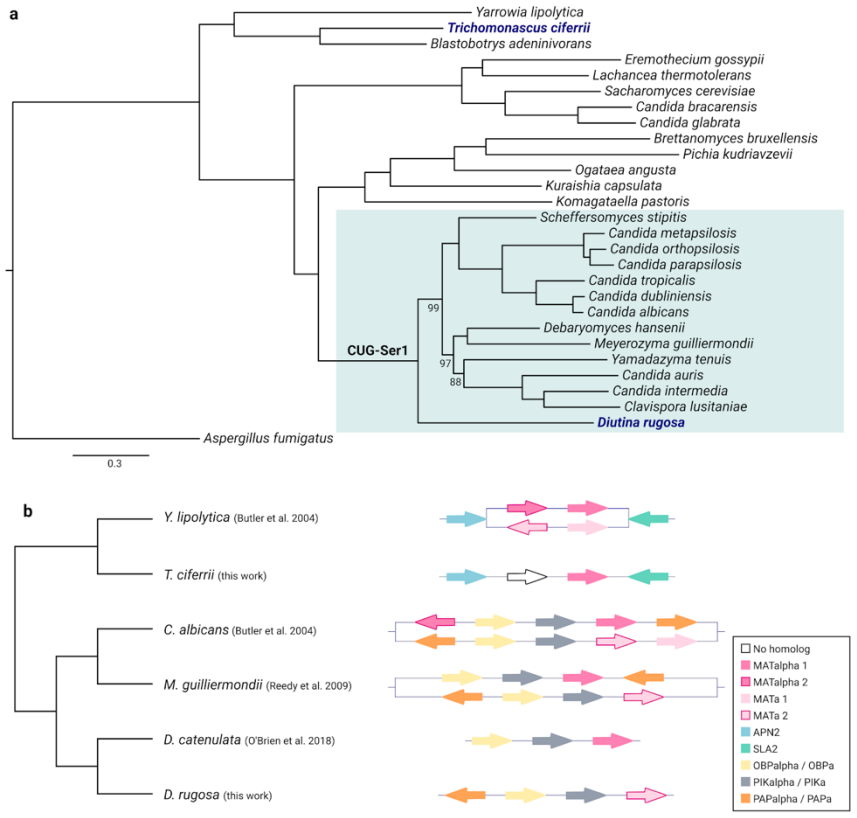

Figure 5_2. Comparative genomics of *D. rugosa* and *T. ciferrii* genomes. **A)** Maximum Likelihood phylogenetic tree of the concatenated alignment of 469 single genes, comprising 297,788 amino-acid positions. When the branch support is different from 100 the values are presented close to the respective branch. CUG-Ser1 clade is highlighted in blue. *D. rugosa* and *T. ciferrii* are marked in dark blue and bold. **B)** Schematic representation of the *MAT* locus of *D. rugosa* and *T. ciferrii* in comparison with closely related species. The tree presents their phylogenetic relationship, but the branch length does not correspond to their phylogenetic distance. Each arrow represents a different gene with the color indicating the gene name.

5.5 Concluding remarks

We have here reported the genomes of two emergent yeast pathogens, which are phylogenetically very distantly related, namely *D. rugosa* and *T. ciferrii*. These two reference genomes provide an important resource for the assessment of relevant aspects of these yeasts, including the genetic bases of their clinically relevant traits, as virulence and drug resistance. In addition, the two phylomes, which include a full repertoire of gene evolutionary histories and a catalog of orthologs and paralogs, can be used to trace the origin and evolution of genes of interest. Therefore, the data provided by this publication will certainly be of interest for the study of emergent yeast pathogens.

5.6 Acknowledgements

This work received funding from the European Union's Horizon 2020 Research and Innovation Program under the Marie Skłodowska- Curie Grant Agreement No. H2020-MSCA-ITN-2014-642095. TG group also acknowledges support from the Spanish Ministry of Economy, Industry, and Competitiveness (MEIC) for the EMBL partnership, and grants 'Centro de Excelencia Severo Ochoa 2013– 2017' SEV-2012-0208, and BFU2015-67107 co-founded by European Regional Development Fund (ERDF); from the CERCA Program/ Generalitat de Catalunya; from the Catalan Research Agency (AGAUR) SGR857, and grants from the European Union's Horizon 2020 Research and Innovation Program under the Grant Agree- ments No. ERC-2016-724173, and MSCA-747607. TG also receives support from an INB grant (PT17/0009/0023 – ISCIII-SGEFI/ ERDF). CLF received funding from Christian Doppler Laboratory for Fungal Infections: Avoid, find, and treat! The authors thank all Gabaldón lab members for helpful discussions on this study, specially Marina Marcet-Houben.

6 Whole-Genome Sequencing of the Opportunistic Yeast Pathogen *Candida inconspicua* Uncovers Its Hybrid Origin

Mixão, V., Hansen, A. P., Saus, E., Boekhout, T., Lass-Flörl, C., & Gabaldón, T. (2019). Whole-Genome Sequencing of the Opportunistic Yeast Pathogen *Candida inconspicua* Uncovers Its Hybrid Origin. *Frontiers in Genetics*, *10*, 383. doi: 10.3389/fgene.2019.00383

6 Whole-Genome Sequencing of the Opportunistic Yeast Pathogen *Candida inconspicua* Uncovers Its Hybrid Origin

6.1 Abstract

Fungal infections such as those caused by *Candida* species are increasingly common complications in immunocompromised patients. The list of causative agents of candidiasis is growing and comprises a set of emerging species whose relative global incidence is rare but recurrent. This is the case of *Candida inconspicua*, which prevalence has increased 10-fold over the last years. To gain novel insights into the emergence of this opportunistic pathogen and its genetic diversity, we performed whole genome sequencing of the type strain (CBS180), and of 10 other clinical isolates. Our results revealed high levels of genetic heterozygosity structured in non-homogeneous patterns, which are indicative of a hybrid genome shaped by events of loss of heterozygosity (LOH). All analyzed strains were hybrids and could be clustered into two distinct clades. We found large variability across strains in terms of ploidy, patterns of LOH, and mitochondrial genome heterogeneity that suggest potential admixture between hybrids. Altogether, our results identify a new hybrid species with virulence potential toward humans and underscore the potential role of hybridization in the emergence of novel pathogenic lineages.

Keywords: *Candida inconspicua*, hybrid, yeast, pathogen, genome

6.2 Introduction

Fungal infections are an increasingly common problem in hospital environments, very often leading to patient's death (Brown et al., 2012; Gabaldón & Carreté, 2016; Lass-Flörl, 2009; Pfaller & Diekema, 2007). Historically, *Candida* species have been the most common causative agents of hospital-acquired fungal infections (Brown et al., 2012; Gabaldón & Carreté, 2016; Lass-Flörl, 2009; Pfaller & Diekema, 2007). Patients at particular risk include those in the intensive care unit, those undergoing surgery, and patients with solid tumor or hematological malignancy (Lass-Flörl, 2009). Reported mortality rates for candidemia range from 28 to 59% in European surveys and depend on species, underlying disease conditions, and geographical location (Jordà-Marcos et al., 2007; Lass-Flörl, 2009). *Candida albicans* is the most common cause of candidemia, accounting for more than 50% of the cases, followed by *Candida glabrata* and *Candida parapsilosis* (Holzheimer & Dralle, 2002; Jordà-Marcos et al., 2007; Pfaller & Diekema, 2007). However, the epidemiology of candidemia has shifted in recent years, with the incidence of rare species becoming increasingly important in the clinical setting (Bretagne et al., 2017; Gabaldón & Carreté, 2016; Gabaldón et al., 2016; Lass-Flörl, 2009; Pfaller et al., 2012; Sardi et al., 2013). For instance, *Candida auris* is an emerging multi-drug resistant pathogen responsible for many outbreaks all over the world in the last few years (Forsberg et al., 2019).

Candida inconspicua was firstly described as *Torulopsis inconspicua* in Lodder & Kreger-van Rij (1952) and later reclassified in *Candida* (Yarrow & Meyer, 1978). The species belongs to the *Pichia cactophila* clade, together with *Pichia kudriavzevii* [synonym *Candida krusei* (Douglass et al., 2018)], *Pichia norvegensis, P. cactophila*, and *Pichia pseudocactophila* (Kurtzman et al., 2011). *C. inconspicua* is genetically similar and phenotypically identical to *P. cactophila* and it has been suggested that they represent different sexual stages of the same species (Guitard et al., 2015; Kurtzman et al., 2008; Kurtzman et al., 2011). *C. inconspicua* has also been misidentified as other members of the clade such as *P. norvegensis* (Guitard et al., 2015; Majoros et al., 2003). Many studies support that *C. inconspicua* can often be found in lactic products, including milk, cheese, or butter (Callon et al., 2007; Krukowski et al., 2006;

Kurtzman et al., 2011; Minervini et al., 2001; Suzzi et al., 2003). In addition, it has recently been reported in traditional alcoholic beverages such as oil palm wine and a sorghum beer called *tchapalo* (Egue et al., 2018).

Candida inconspicua is also responsible for clinical infections, more prominently in European countries (Guitard et al., 2013; Majoros et al., 2005; Pfaller et al., 2010b). A more than 10-fold increase in *C. inconspicua* infections between 1997–2000 and 2001–2004 (increase of 9 to 276 cases) followed by an apparent stabilization has been reported by a multi-center study (Pfaller et al., 2010b). The majority of *C. inconspicua* infections are associated with osteomyelitis, oropharyngeal and esophageal candidiasis in HIV positive patients, as well as with candidemia in patients with hematological malignancies (Majoros et al., 2005). Frequently, *C. inconspicua* isolates derive from colonization of the digestive and respiratory tracts from unknown sources (Guitard et al., 2013). However, the above-mentioned reported isolations make contaminated milk or other food products a possible source for the infecting strains. *C. inconspicua* was previously described as presenting a low susceptibility to fluconazole and other antifungal agents (Guitard et al., 2013; Majoros et al., 2005; Pfaller et al., 2010b). For instance, Pfaller et al., (2010b) reported that, depending on the site of isolation, the frequency of fluconazole resistant strains could range between 26.1% (skin and soft tissue) and 62.9% (genital tract), thus indicating a high phenotypic heterogeneity among *C. inconspicua* isolates.

To shed light on the genetic makeup and diversity of this emerging opportunistic pathogen we undertook the whole genome sequencing and assembly of the type strain and compared the genomic sequences of 10 other clinical isolates. We found that *C. inconspicua* has a highly heterozygous genome with patterns suggestive of a hybrid origin. We discuss this finding in comparison with two other medically important *Candida* hybrid lineages: *Candida metapsilosis* and *Candida orthopsilosis* (Pryszcz et al., 2014, 2015; Schröder et al., 2016). Following *C. metapsilosis*, *C. inconspicua* is the second reported case of an opportunistic *Candida* human pathogen for which all clinical strains analyzed so far are hybrids, suggesting that hybridization may be at the root of its ability to infect humans (Mixão & Gabaldón, 2018).

6.3 Materials and Methods

Library Preparation and Genomic DNA Sequencing

For this project we sequenced *C. inconspicua* type strain (CBS180), and 10 other clinical isolates. These isolates were sent by collaborating laboratories in the frame of an antifungal susceptibility test project. They were grown on Sabouraud's 2% dextrose agar at 30°C for 72 h and re-identified using direct extraction method by MALDI-TOF (MALDI-Biotyper, Bruker, Daltonics, Database version, United States). Genomic DNA extraction of the 11 *C. inconspicua* strains was performed using the MasterPure Yeast DNA Purification Kit (Epicentre, United States) following manufacturer's instructions. Briefly, *C. inconspicua* cultures were grown in an orbital shaker overnight (200 rpm, 30°C) in 15 ml of YPD medium. Cells were harvested using 4.5 ml of each culture by centrifugation at maximum speed for 2 min, and then they were lysed at 65°C for 15 min with 300 µl of yeast cell lysis solution (containing 1 µl of RNAse A). After being on ice for 5 min, 150 µl of MPC protein precipitation reagent were added into the samples, and they were centrifuged at 16.000 *g* for 10 min to pellet the cellular debris. The supernatant was transferred to a new tube, DNA was precipitated using 100% cold ethanol and centrifuging the samples at 16.000 *g*, 30 min, 4°C. The pellet was washed twice with 70% cold ethanol and, once the pellet was dried, the sample was resuspended in 100 µl of TE. All gDNA samples were cleaned to remove the remaining RNA using the Genomic DNA Clean & Concentrator kit (Epicentre) according to manufacturer's instructions. Total DNA integrity and quantity of the samples were assessed by means of agarose gel, NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, United States) and Qubit dsDNA BR assay kit (Thermo Fisher Scientific).

Whole-genome sequencing was performed at the Genomics Unit from Centre for Genomic Regulation (CRG) with a HiSeq2500 machine. Libraries were prepared using the NEBNext Ultra DNA Library Prep kit for Illumina (New England BioLabs, United States) according to manufacturer's instructions. All reagents subsequently mentioned are from the NEBNext Ultra DNA Library Prep kit for Illumina if not specified otherwise. 1 µg of gDNA was fragmented

by nebulization in Covaris to a size of ~600 bp. After shearing, the ends of the DNA fragments were blunted with the End Prep Enzyme Mix, and then NEBNext Adaptors for Illumina were ligated using the Blunt/TA Ligase Master Mix. The adaptor-ligated DNA was cleaned-up using the MinElute PCR Purification kit (Qiagen, Germany) and a further size selection step was performed using an agarose gel. Size-selected DNA was then purified using the QIAgen Gel Extraction Kit with MinElute columns (Qiagen) and library amplification was performed by PCR with the NEBNext Q5 Hot Start 2X PCR Master Mix and index primers (12–15 cycles). A further purification step was done using AMPure XP Beads (Agentcourt, United States). Final libraries were analyzed using Agilent DNA 1000 chip (Agilent) to estimate the quantity and check size distribution, and they were then quantified by qPCR using the KAPA Library Quantification Kit (KapaBiosystems, United States) prior to amplification with Illumina's cBot. Libraries were loaded and sequenced 2 × 125 bp on Illumina's HiSeq 2500. Base calling was performed using Illumina pipeline software. In multiplexed libraries, we used 6 bp internal indexes (5′ indexed sequences). De-convolution was performed using the CASAVA software (Illumina, United States). Sequence data of the genomes has been deposited in short read archive (SRA) under the BioProject Accession No. PRJNA517794.

De novo Genome Assembly and Read Mapping

Raw sequencing data was inspected with FastQC v0.11.5. Paired-end reads were filtered for quality below 10 or size below 31 bp and for the presence of adapters with Trimmomatic v0.36 (Bolger et al., 2014). The K-mer Analysis Toolkit (KAT; Mapleson et al., 2017) was used to get the GC content and *k*-mer frequency of CBS180 reads and estimate the expected genome size. SPAdes v3.9 (Bankevich et al., 2012) was used to perform the genome assembly using this strain. Afterward, redundant contigs were removed with Redundans v0.13c (Pryszcz & Gabaldón, 2016) using default parameters, i.e., 51% minimum identity and at least 80% overlap. The quality of the assembly was inspected with Quast v4.5 (Gurevich et al., 2013) and KAT (Mapleson et al., 2017). Genome annotation was performed with Augustus v3.1 (Stanke & Morgenstern, 2005), using *C. albicans* as model organism. BUSCO v3 (Waterhouse et al., 2018) was used to assess the completeness predicted proteome considering the

Ascomycota database. This genome assembly and annotation have been deposited at DDBJ/ENA/GenBank under the Accession No. SELW00000000. The version described in this paper is version SELW01000000.

Phylome reconstruction was performed using the PhylomeDB pipeline (Huerta-Cepas et al., 2014) as described by Pryszcz et al., (2015), using the predicted proteome as seed, and considering other twenty-one Saccharomycotina species (Table A in Supplementary file 1). A second phylome considering only proteins predicted in scaffolds > 10 kb was also reconstructed to confirm the obtained results. The presented results correspond to the phylome considering all proteins. *C. inconspicua* phylomes are available in PhylomeDB (Huerta-Cepas et al., 2014) with the ID 454 and 498 (this one only considering scaffolds > 10 kb). Gene gain and loss analysis in seed branch was performed based on the phylome results. A BLASTp (Zhang et al., 2000) was performed against the UniProt database, in order to determine the possible function associated to these genes, as well as their GO terms. An enrichment analysis was done using FatiGO (Al-Shahrour et al., 2004), and the results were summarized with REVIGO (Supek et al., 2011).

Read mapping for all strains (Table 6_1) was performed with BWA-MEM v0.7.15 (Li, 2013). Picard v2.1.1 was used to sort the resulting file by coordinate, as well as to mark duplicates, create the index file, and obtain mapping statistics. Mapping results were inspected with IGV version 2.0.30 (Thorvaldsdóttir et al., 2013). Mapping coverage was determined with SAMtools v0.1.18 (Li et al., 2009).

Variant Calling and Ploidy Estimation

Samtools v0.1.18 (Li et al., 2009) and Picard v2.1.12 were used, respectively, to index the reference and create a dictionary to be used in subsequent variant calling steps. GATK v3.6 (McKenna et al., 2010) was used to call variants with the tool HaplotypeCaller set with –genotyping_mode DISCOVERY -stand_emit_conf 10 -stand_call_conf 30 -ploidy 2 -nct 8. The tool VariantFiltration of the same program was used to filter the vcf files with the following parameters: –clusterSize 5 –clusterWindowSize 20 –genotypeFilterName “heterozygous” –genotypeFilterExpression

"isHet = =1" –filterName "bad_quality" -filter "QD < 2.0 || MQ < 40 || FS > 60.0 || HaplotypeScore > 13.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0" –filterExpression "DP < = 20" –filterName "DepthofQuality." In order to determine the number of SNPs/kb, a file containing only SNPs was generated with the SelectVariants tool. Moreover, for this calculation only positions in the reference with 20 or more reads were considered for the genome size, and these were determined with bedtools genomecov v2.25.0 (Quinlan & Hall, 2010).

To estimate the ploidy of each strain, nQuire histotest (Weiß et al., 2018) was used to test which distribution (diploid, triploid, or tetraploid) fits better to the variant frequency data. Given that for some of the strains the results were not clear, the allele frequency of each heterozygous variant was calculated by dividing the number of reads supporting the alternative haplotype by the total of reads mapping at that position. Allele frequency density was plot for each strain considering only scaffolds with more than 100 kb.

Determination of the Different Hybridization Events and Parental Divergence

To determine heterozygous and loss of heterozygosity (LOH) blocks, the procedure applied and validated by Pryszcz et al., (2015) was used. Briefly, bedtools merge v2.25.0 (Quinlan & Hall, 2010) with a window of 100 bp was used to define heterozygous regions, and by opposite, LOH blocks would be all non-heterozygous regions in the genome. The minimum LOH and heterozygous block size was established at 100 bp. Due to the aneuploidies observed for *C. inconspicua*, contrary to what Pryszcz and colleagues performed, no coverage filter was applied.

As mentioned in section "Results", all *C. inconspicua* strains analyzed here were found to be hybrids. The current divergence between the parental genomes was calculated dividing the number of heterozygous positions by the total size of heterozygous blocks. Another important step was to check whether all strains were originated from the same hybridization event. For this, pairwise comparisons between overlapping LOH blocks were performed using bedtools jaccard v2.25.0 (Quinlan & Hall, 2010), which

allowed us to get the number of nucleotides in the intersection of the two strains over the number of nucleotides present in their union. Moreover, assuming that LOH blocks with exactly the same boundaries are not independent events, it was decided to repeat the pairwise comparisons, but this time instead of analyzing the number of nucleotides in LOH regions in both strains, it was decided to get the number of LOH blocks with exactly the same boundaries in both strains. In this case, in order to avoid false positives, short LOH blocks (<1 kb), as well as very short scaffolds (<10 kb) were not considered for the analysis.

Phylogenetic Analysis

To obtain the phylogenetic relationship between the 11 strains, FastaAlternateReferenceMaker tool of GATK v3.6 (McKenna et al., 2010) was used to obtain the genome sequence of each strain substituting each position with a homozygous SNP by the respective allele. Furthermore, bedtools subtract and bedtools getfasta v2.25.0 (Quinlan & Hall, 2010) were used to remove from these new sequences all positions with a heterozygous SNP or an INDEL in at least one of the strains. In the end, as INDELs were not considered for the analysis, the sequences of the 11 strains presented exactly the same size, constituting a sequence alignment of 9,971,439 bp. A Maximum-likelihood tree representative of this alignment was obtained with RAxML v8.2.8 software (Stamatakis, 2014), using the GTRCAT model. The same approach was applied to obtain a phylogeny for MAT locus.

Lineage Prediction and Detection of Recombination

To predict the number of lineages and clusters in our dataset, as well as to detect traces of admixture between the different strains, we used fastGEAR (Mostowy et al., 2017). For that, this program was set to complete mode, and the same alignment used for the phylogenetic analysis was given as input. All the other parameters were set to default. To have a control of the two expected scenarios, we decided to do the same analysis for *C. metapsilosis* and *C. orthopsilosis*, as representatives of a unique hybridization event and multiple hybridization events, respectively (Pryszcz et al., 2015; Schröder et

al., 2016). In both situations, all Illumina paired-end sequencing libraries available (BioProjects PRJEB4430, PRJEB1698 and PRJNA322245), as well as five extra libraries for *C. metapsilosis* and two other libraries for *C. orthopsilosis*, which will be soon publicly available under the BioProject PRJNA520893 (manuscript in preparation) were used for read mapping, post-processing analysis, variant calling and sequence alignment as mentioned before for *C. inconspicua*.

Mitochondrial Genome Assembly

The mitochondrial genome assembly for *C. inconspicua* was performed using the filtered Illumina paired-end reads of CI1 strain. NOVOPlasty v2.7.2 (Dierckxsens et al., 2017) with default parameters was used to assemble this genome, taking as seed input *C. inconspicua Cox2* gene (Accession No. EF599394.1). A final 31 kb contig was obtained. NUCmer algorithm of MUMmer v3.1 (Kurtz et al., 2004) was used to align this final assembly against *Pichia kluyveri* mitochondrial genome (Accession No. NC_022158.1). MUMmerplot algorithm of MUMmer v3.1 (Kurtz et al., 2004) was used to visualize this alignment and see that our final 31 kb scaffold covers a big part of *P. kluyveri* mitochondrial genome (Supplementary Figure 1). Mitochondrial genome annotation was performed with MITOS2 (Bernt et al., 2013). Read mapping and variant calling of all strains against this final mitochondrial assembly was performed as mentioned before. The mitochondrial genome of each strain was reconstructed with FastaAlternateReferenceMaker tool of GATK v3.6 (McKenna et al., 2010), using IUPAC code to solve heterozygous positions. A NJ phylogenetic tree was generated with SplitsTree v4 (Huson & Bryant, 2006). To compare the topology of mitochondrial and nuclear trees, RAxML v8.2.8 software (Stamatakis, 2014) was used to compute per site log Likelihoods for each tree given each of the alignments. Consel v1.2 (Shimodaira & Hasegawa, 2001) was used to assess the confidence that a given tree could correspond to a given alignment.

Antifungal Susceptibility Test

To test whether the different clades of *C. inconspicua* presented different susceptibilities to antifungal drugs, we performed antifungal susceptibility tests on all 11 strains using two different methods (Etest gradient strips and EUCAST). By EUCAST broth-microdilution, which is one of the main international reference methods, the following drugs were tested: Itraconazole (Sigma, Rowville, Australia), Posaconazole (Schering-Plough, Kenilworth, NJ, United States), Isavuconazole (Basilea, Basel, Switzerland), Fluconazole (Sigma), Voriconazole (Sigma), Anidulafungin (Pfizer, New York, NY, United States), Micafungin (Astellas, Munich, Germany), Caspofungin (Sigma), and Amphotericin B (Sigma) in Cellstar plates (Cellstar Cat-No. 655180, Greiner Bio-One, United States). Pre-cultures were grown on Sabouraud's 2% dextrose agar at 30°C for 24 h for all method used. The RPMI used for the different media was provided by Sigma (RPMI-1640 Medium, R6504-50L). Broth-microdilution was performed according to EUCAST guidelines (Arendrup et al., 2017) with minor modifications. To ensure proper growth, the incubation time was prolonged to 48 h, and the optical density threshold of the plate reader reading was lowered to 0.1. Plates were evaluated at 48 h both visually and by plate reader (Microplate Reader model 680, Bio-Rad, United States). *Candida parapsilosis* ATCC 22019 or *P. kudriavzevii* ATCC 6258 were used as quality control. On the other hand, a commercial test was also used. Specifically, Etest strips with Itraconazole, Posaconazole, Isavuconazole, Fluconazole, Voriconazole, Anidulafungin, Micafungin, Caspofungin, and Amphotericin B (all bioMérieux SA, France, except Isavuconazole, which was provided by Liofilchem, Italy) were used as indicated by the respective manufacturers. Plates were incubated at 37°C and visually read after 48 h to coincide with the EUCAST conditions.

6.4 Results

Evidence for the Hybrid Nature of *C. inconspicua*

To uncover the genomic features of *C. inconspicua* we sequenced the type strain CBS180 using an Illumina-based, pair-end sequencing strategy (see section "Materials and Methods"). GC content and 27-mer count analyses of the sequencing reads revealed two peaks with similar GC content but different coverage (Figure 6_1A, B). The first peak presented roughly half of the coverage of the second (Figure 6_1B), and therefore it could correspond to highly heterozygous regions of a diploid genome. We next performed a *de novo* genome assembly using SPAdes (Bankevich et al., 2012) and redundans (Pryszcz & Gabaldón, 2016), an assembly pipeline tailored for highly heterozygous genomes (see section "Materials and Methods"). The final assembly comprised 10,353,411 bp divided in 744 contigs (76 longer than 50 kb, representing 80.16% of the genome) with 35.1% GC content and a N50 of 100,257 bp. Genome annotation predicted 5,079 proteins (see section "Materials and Methods"). The final genome size and the number of proteins that we have obtained are similar to what was previously described for the closely related species *P. kudriavzevii* (10.9 Mb and 4,949 predicted proteins; Cuomo et al., 2017), suggesting the completeness of the *C. inconspicua* genome. Indeed, 99.13% of CBS180 reads aligned to the assembly, and the missing 27-mers were in heterozygous regions, possibly corresponding to redundant contigs (Figure 6_1B). Furthermore, our predicted proteome has a 89% completeness as assessed by BUSCO (see section "Materials and Methods"). Phylome reconstruction (Gabaldón, 2008) in the context of twenty-one other Saccharomycotina species (see section "Materials and Methods" and Table A in Supplementary File 1) identified 501 species-specific genes. Furthermore, genes specifically duplicated in *C. inconspicua* seemed to be enriched in transmembrane transport and drug export functions, among others (Table B in Supplementary File 1).

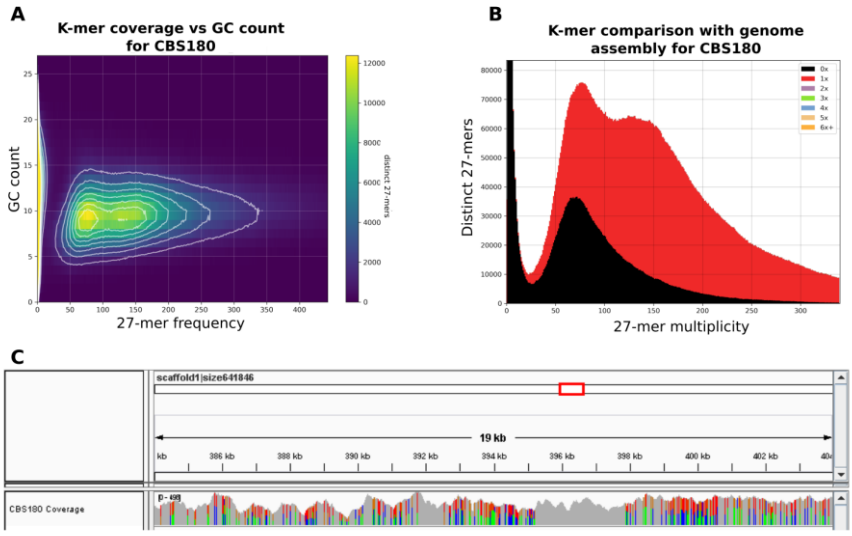

Figure 6_1. Heterozygosity patterns in *Candida inconspicua* type strain genome. **(A)** 27-mer frequency in CBS180 genomic reads and respective GC content. **(B)** 27-mer frequency in CBS180 genomic reads and indication of their presence (red) or absence (black) in the genome assembly. Although both 27-mer count peaks were represented in the genome assembly, roughly half of the reads corresponding to the first one was excluded. These reads probably correspond to redundant heterozygous contigs that were intentionally removed during the reduction step of the assembly process (see section "Materials and Methods"). **(C)** IGV coverage track of a 19 kb region of *C. inconspicua* scaffold 1. Colors represent polymorphic positions. Highly heterozygous regions are clearly interspaced by blocks of loss of heterozygosity (LOH).

Mapping of CBS180 reads against its own genome assembly followed by variant calling showed the presence of 14.36 variants/kb, of which most (14.34 variants/kb) corresponded to heterozygous positions (Table 6_1 and Supplementary File 2). Importantly, these variants were not homogeneously distributed throughout the genome, but rather formed stretches of highly heterozygous sequences separated by what appeared to be LOH, as it has been observed in previously analyzed genomes from hybrids of the *Candida* clade (Pryszcz et al., 2014, 2015; Schröder et al., 2016). These LOH blocks were flanked by heterozygous blocks with relatively constant levels of heterozygosity (36.4 heterozygous variants/kb). An illustrative example of such patterns is presented in Figure 6_1C. The high level of sequence divergence between the two haplotypes in the

heterozygous blocks (3.64%) is much higher than the divergence observed between most distantly related strains of well-recognized yeast species (i.e., 1.1% for *Saccharomyces cerevisiae*) (Peter et al., 2018). Altogether, these analyses highly suggested that *C. inconspicua* type strain is a hybrid with a chimeric genome.

Genome Heterogeneity Across *C. inconspicua* Strains

To gain a better insight into this species, 10 other clinical strains were sequenced, and their sequencing reads were mapped against the reference genome assembly described above (Table 6_1). All analyzed strains were highly heterozygous, with 14.15–19.76 heterozygous variants/kb, and all of them presented highly heterozygous genomic regions interspersed with LOH blocks (Table 6_1, Supplementary File 2 and Supplementary Figure 2). Thus, similar to the previously described hybrids in *C. metapsilosis* and *C. orthopsilosis* (Pryszcz et al., 2014, 2015; Schröder et al., 2016), *C. inconspicua* clinical isolates seemed to comprise a majority of hybrid strains, with all 11 strains analyzed so far being hybrids. The presence of both **a** and **alpha** mating-types in the *MAT* locus suggests matting as a possible origin of the hybrids (Supplementary Figure 3).

Similar to previous studies (Pryszcz et al., 2014), the non-homogeneous distribution of heterozygous variants throughout the genome allowed us to define blocks of heterozygosity (see section "Materials and Methods"), which correspond to regions that retain genetic material from both hybridized lineages. On average, each strain presented 12,044 heterozygous blocks with an average size of 326 bp each, overall covering 38.15% of the genome, and comprising 82.41% of the heterozygous variants (Supplementary File 2). Based on the density of heterozygous variants within heterozygous blocks, we estimated that the current sequence divergence at the nucleotide level between the two parental lineages is approximately 3.72% (Figure 6_2).

Table 6_1. List of *C. inconspicua* strains used in this project, with indication of their respective clade, place of collection, specimen, number of heterozygous variants, level of loss of heterozygosity and estimated overall ploidy. Non-type strains were collected in the last 10 years in the framework of antifungal susceptibility testing.

Strain	Clade	Country	Specimen	Heter variants / kb	Estimated LOH >100bp	nQuire most probable ploidy
14ANR23920	Clade 1	Germany	Blood	14.66	65.18%	Diploid
9_16	Clade 1	Belgium	Blood	14.57	65.43%	Diploid
CI1	Clade 1	Austria	Abdominal fluid	14.15	66.45%	Diploid
CBS180*	Clade 2	Netherlands	Sputum	14.34	65.11%	Diploid
110_10	Clade 2	Austria	Blood	18.18	55.98%	Triploid
1282	Clade 2	Romania	Swab	18.88	54.06%	Triploid
CNM_CL6867	Clade 2	Spain	Swab	19.35	53.01%	Triploid
IUM_96-0030	Clade 2	Italy	Swab	18.84	54.21%	Triploid
LL867	Clade 2	Spain	Blood	19.76	51.94%	Triploid
NRZ_BK_345	Clade 2	Germany	Blood	19.70	52.78%	Triploid
UCSC_1590	Clade 2	Italy	Blood	19.30	53.06%	Triploid

*Type strain

We next used the called SNPs to reconstruct phylogenetic relationships between the sequenced *C. inconspicua* strains using a maximum likelihood approach (see section "Materials and Methods"). The resulting strain phylogeny revealed the presence of at least two distinct clades (Figure 6_3), with the strains 14ANR23920, CI1, and 9_16 forming one clade (clade 1) separated by a long branch from another clade comprising the remaining strains (clade 2). Within clade 2, IUM_96-0030 appeared at a relatively long distance from the rest of the clade, as did, to a lesser extent, CBS180 and 1282 (Figure 6_3). The two clades were not related with the geographical distribution of the different strains, but invasive strains seemed to form two clusters, one in each of the clades (Figure 6_3 and Table 6_1). Furthermore, the levels of susceptibility of each strain to azoles, echinocandins, or amphotericin B (see section "Materials and Methods") seemed to be unrelated with their position in the phylogeny (Supplementary File 3).

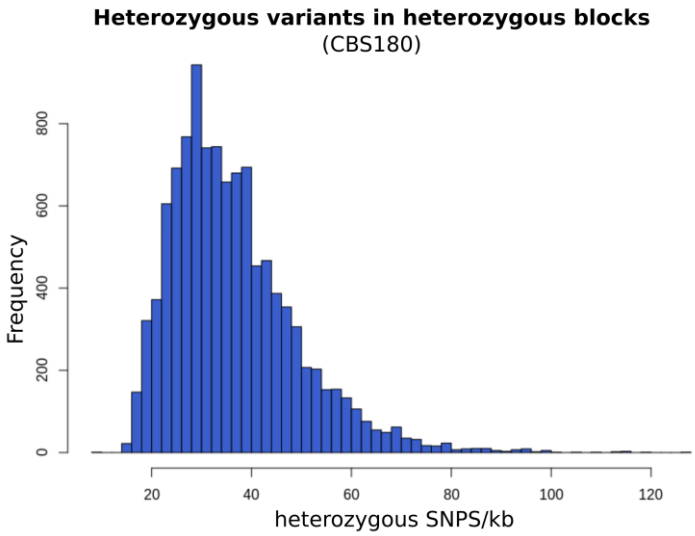

Figure 6_2. Frequency of heterozygous variants per kilo-base in CBS180 heterozygous blocks. The distribution of this frequency is close to normal, with a peak at 30 variants/kb, consistently with the estimated current parental divergence.

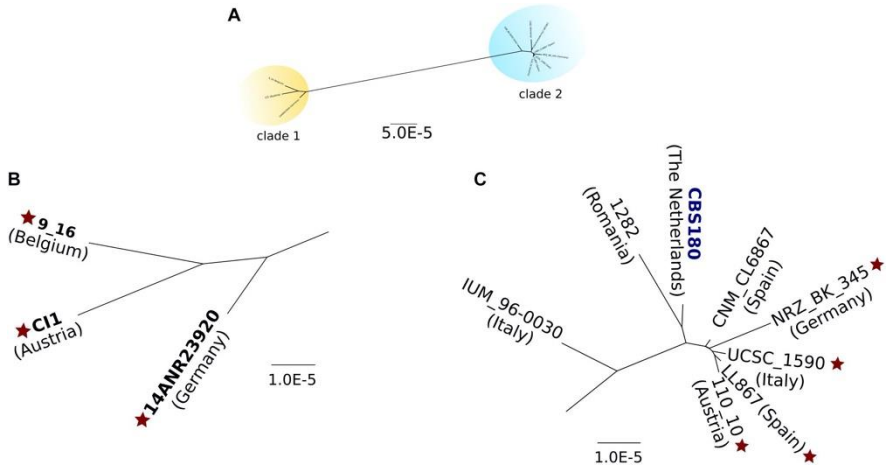

Figure 6_3. Maximum likelihood phylogeny of all *C. inconspicua* strains. A concatenated alignment of 9,971,439 bp was used to reconstruct this phylogeny. Positions with heterozygous variants or INDELs in at least one strain were removed from the analysis. The type strain is marked in dark blue. Diploid strains are highlighted in bold. Star indicates invasive strains. **(A)** Overall view of the nuclear genome phylogeny of *C. inconspicua*, with two distinct clades highlighted in yellow (clade 1) and blue (clade 2). **(B)** Closer view on *C. inconspicua* clade 1. **(C)** Closer view on *C. inconspicua* clade 2.

The existence of two separate hybrid clades pointed to two possible scenarios: (i) a unique hybridization event followed by ancestral divergence of the two clades; or (ii) two independent hybridization events between the same parental lineages, each originating one of the clades. To explore these scenarios, we first analyzed the *MAT* locus of the different strains (Supplementary Figure 3). While for mating-type alpha we could identify two major clades, which coincide with the nuclear genome phylogeny, for mating-type **a**, a third clade formed by CBS180 and IUM_96-0030 was detected (Supplementary Figure 3). Thus, this analysis did not allow us to clearly support one or the other scenario. In a second attempt to explore these scenarios, we used a previously described approach based on the comparison of patterns of LOH blocks between strains (Pryszcz et al., 2015), where the presence of large LOH blocks with similar boundaries is indicative of a common origin. To do so, we defined LOH blocks for each *C. inconspicua* strain (see section "Materials and Methods"). On average, each strain presented 20,620 blocks with an average size of 292 bp, which covered 57.93% of the genome. Expectedly, pairwise comparisons of the overlap between LOH of each two strains revealed the same two clades as the phylogenetic analysis, but with CBS180, 1282 and IUM_96-0030 not being clearly classified to any of them (Supplementary Figure 4). To investigate whether the two clades resulted from two different hybridization events we identified LOH blocks with exactly the same boundaries in each pairwise comparison of the strains. Considering only blocks at least 1 kb-long present in scaffolds longer than 10 kb resulted in the same two clades (Supplementary File 2). However, all pairwise comparisons shared at least five large LOH blocks with exact boundaries. Indeed, disregarding any size limit in LOH blocks or scaffold size, all strains shared 633 blocks (3.07% of the average number of blocks, Supplementary Figure 5 as example). The relative low number of shared blocks did not allow us to clearly discard the possibility that more than one hybridization event have occurred. In fact, cluster and lineage prediction (see section "Materials and Methods"), using *C. metapsilosis* and *C. orthopsilosis* as control scenarios for one and more than one hybridization events, respectively, clearly identified two different lineages in *C. inconspicua* (Supplementary Figure 6). Therefore, we considered that the most probable scenario is that all these strains were originated by two hybridization events, each one generating one of the clades.

Aneuploidies and Mitochondrial Genome Heterogeneity Suggest Possible Admixture Among *C. inconspicua* Hybrids

We next estimated the ploidy level of each *C. inconspicua* strain using patterns of allele frequency (see section "Materials and Methods"). Our results pointed to the existence of two different groups of strains. Firstly, CBS180, and the three strains of clade 1, 14ANR23920, 9_16, and CI1, were mainly diploid ($r^2 > 0.9$, Table 6_1). In contrast, the remaining seven strains presented a triploid model as the one with best support, although the results were inconclusive (Supplementary File 4). A closer inspection of allele frequency plots for individual scaffolds suggested the existence of a large number of aneuploidies, as shown by the coexistence of patterns compatible with tetraploid, triploid and diploid configurations, or a mixture thereof, in different scaffolds of a given strain (Supplementary Figure 7). The number of such aneuploidies was very reduced in the first group of strains with a largely diploid structure. All strains with large levels of aneuploidies belonged to the phylogenetic clade 2. In this regard, the type strain (CBS180) was the only diploid strain within a clade consisting mostly of aneuploid strains. This anomaly was perhaps related to the fact that the type strain has been conserved in isolated culture for a long time, likely promoting a fast diploidization, due to its frequent subculturing. Aneuploidies in the other strains may indicate intermediate levels of diploidization from an allotetraploid parental.

In another attempt to clarify the origin of these hybrid strains, we assembled a 31 kb region of *C. inconspicua* mitochondrial genome (see section "Materials and Methods"). This region was obtained from CI1 strain (clade 1), as with the type strain (CBS180) we could only get a highly fragmented mitochondrial assembly (see section "Materials and Methods"). The reads of all strains were mapped against this region. Overall, the patterns of variation in the mitochondrial genome were consistent with the nuclear genome phylogeny (Figure 6_3 and Supplementary Figures 8 and 9), with most polymorphisms likely resulting from accumulation of SNPs through time in the same mitochondrial genome background, rather than representing two different mitochondrial genomes each coming from a different parental species of the hybrids. Interestingly, mitochondrial genomes from all strains of clade 2 revealed some

short deleted regions, from which we highlight a major 1.5 kb deletion in *Cox1* (Supplementary Figure 8), corresponding to a LAGLIDADG endonuclease domain, which is a mobile element.

As expected for a scenario of hybridization, the phylogenetic tree of the mitochondrial genome (Supplementary Figure 9) did not present the same topology as the one of the nuclear genome (AU test p-val < 0.033), Even so, we could identify exactly the same clades and sub-clades in the two trees (Supplementary Figure 9). Importantly, the IUM_96-0030 strain showed heterogeneous variation patterns consistent with the presence of mitochondrial sequences from the two clades (Supplementary Figure 8). For instance, besides the presence of heterozygous SNPs in the mitochondrial sequence, this strain presented some coverage in the above mentioned 1.5 kb deletion, indicating that this strain presented heteroplasmy (i.e., presence of two mitochondrial sequences within the same cell). This suggested that this clade 2 strain may have fused with a *C. inconspicua* strain belonging to clade 1. Indeed, fastGEAR (Mostowy et al., 2017) identified recent recombination in some nuclear regions of IUM_96-0030 whose source was clade 1 (Supplementary File 5). The unbalanced representation of the two types of mitochondria in IUM_96-0030 could be related to an unbalance in the inheritance of the mitochondrial genome, which was previously postulated to occur in yeast hybrids (Verspohl et al., 2018). This pointed to the occurrence of recent admixture between different *C. inconspicua* strains. Considering this, a plausible scenario is that the aneuploidies mentioned above, and the mitochondrial genome heterogeneity described here are not unrelated phenomena, and both indicate that several strains of clade 2 result from recent fusions between *C. inconspicua* hybrids.

6.5 Discussion

Candida inconspicua is an opportunistic pathogenic yeast which is increasingly reported as a source of infection and often presents antifungal resistance (Arendrup & Patterson, 2017; Cendejas-Bueno et al., 2010; Guitard et al., 2013; Sugita et al., 2004). In this work, we have *de novo* assembled its type strain (CBS180) and sequenced 10 additional clinical isolates to obtain a better understanding of its

recent evolution. Our results show compelling evidence for a hybrid nature of the *C. inconspicua* lineage, with all strains analyzed so far being hybrids between the same two parental lineages. Sequenced strains clearly clustered into two distinct clades. Although we consider that our results are not sufficiently conclusive, the low frequency of shared LOH blocks, and the identification of two different lineages by fastGEAR (Mostowy et al., 2017) make two independent hybridization events as the most plausible scenario.

All strains analyzed in this work were collected in Europe and the two clades do not show a particular enrichment in any geographical area. Although the inclusion of a larger number of strains may reveal geographical patterns in the future, this pattern is reminiscent of those of *C. metapsilosis*, where only a single hybrid lineage was found to have a global distribution, and *C. orthopsilosis*, where four different clades representing independent hybridization events have been identified but each of which has a widespread distribution (Pryszcz et al., 2015; Schröder et al., 2016).

Hybrid genomes present high levels of heterozygosity which may result in negative epistatic interactions and, consequently, reduced fitness (Mixão & Gabaldón, 2018). Such negative effects of hybridization can sometimes be compensated by emerging phenotypic properties that enable adaptation to a new niche, and therefore may offer a competitive advantage in certain circumstances. In any case, hybridization is generally followed by different processes that lead to a gradual stabilization of these genomes, like whole genome duplication, LOH, or partial or complete chromosome loss (Mixão & Gabaldón, 2018). In *C. inconspicua*, we could observe different aneuploidies, which might represent intermediate ploidy stages to achieve the so-called genome stabilization. For instance, we could identify two groups of strains. A first group, mainly diploid, with less heterozygous variants and a higher level of LOH, and a second one with different ploidy levels, and less LOH. These two groups are almost completely coincident with the two phylogenetic clades, except for the mainly diploid type strain CBS180, which is intermingled in a clade of mostly aneuploid strains (clade 2, Figure 6_3). This clear diploid status of CBS180 is atypical of strains in clade 2. Furthermore, contrary to what is generally expected for hybrid genomes, where the improper chromosome pairing can cause problems during meiosis (Mixão & Gabaldón, 2018), CBS180 is able

to enter meiosis and form ascospores (Guitard et al., 2015). This might be related to the apparent genomic stabilization that we have observed for this strain. Indeed, this strain is the type strain of *C. inconspicua* and therefore is being kept in collections for many years, completely isolated from all the other strains. We cannot be certain of whether this strain was diploid when it was collected or not, but we believe that the isolated environment and the recurrent subculturing and consequent bottleneck to which this strain is subject can have helped the genome stabilization and contributed to its different genomic patterns when compared to all the other strains.

The high prevalence of aneuploidies in some *C. inconspicua* strains is unexpected when compared to the evolutionary patterns observed in other hybrids (Pryszcz et al., 2014, 2015; Schröder et al., 2016). Additionally, the analysis of their mitochondrial genomes suggests the occurrence of crosses between different *C. inconspicua* strains and enabled us to distinguish the same three sub-clades of strains within clade 2 as the nuclear phylogeny (Figure 6_3 and Supplementary Figure 8). For instance, a sub-clade formed exclusively by IUM_96-0030, another one formed by CBS180 and 1282, and a third one with the remaining strains of clade 2. This suggests that clade 2 is formed by three sub-clades of hybrid strains, which are now in a process of diploidization after probably independent crossing events of two *C. inconspicua* strains. A possible scenario is that, in a process similar to the *C. albicans* parasexual cycle (Forche et al., 2008), two diploid hybrid strains form a tetraploid intermediate that would subsequently lose chromosomes until a diploid state is regained. Another possible explanation is that possibly originated ascospores can eventually cross with other *C. inconspicua* strains, working as a source of variability.

6.6 Conclusion

Candida inconspicua is a lineage comprising opportunistic pathogens with a hybrid origin. Although the number of tested strains is low, the absence of homozygous parentals among clinical isolates suggests that the parental lineages are/might be less able to cause infections when compared to hybrid strains. This adds to a growing list of hybrid yeast opportunist lineages and underscores the

relevance of hybridization in the origin of new virulent lineages (Mixão & Gabaldón, 2018). The level of genetic variability among *C. inconspicua* hybrid strains is high, including distinct levels of aneuploidies and the presence of mitochondrial heterogeneity. This suggests that *C. inconspicua* hybrids are plastic and prone to adapt to new environments (Mixão & Gabaldón, 2018). Given the medical importance of this species, this should represent a special concern, as this high genomic plasticity may also correlate to a larger phenotypic diversity and a higher propensity to adapt to antifungal drugs and develop new resistances. Thus, more studies to identify new hybrid pathogens, as well as to try to understand how they shape their genomes, and how they can adapt to new environments should be performed. Indeed, it would be very interesting to analyze the genome of environmental strains, to understand if they are also hybrids, or if the hybridization event was the trigger that made this species become a pathogen, as it is suggested for *C. metapsilosis* (Pryszcz et al., 2015).

6.7 Data availability

Raw sequencing reads generated for this study can be found in SRA under the BioProject Accession No. PRJNA517794. Genome assembly and annotation are available under the same BioProject.

6.8 Author contributions

AH, TB, and CL-F provided the strains and strain information. AH and CL-F performed susceptibility experiments. ES extracted DNA and prepared material for sequencing. VM performed all bioinformatics analyses. TG supervised the study. TG and VM wrote a first draft of the manuscript. All authors contributed to the final manuscript.

6.9 Funding

This work received funding from the European Union's Horizon 2020 Research and Innovation Program under the Marie

Skłodowska-Curie Grant Agreement No. H2020-MSCA-ITN- 2014-642095. TG group also acknowledges support from the Spanish Ministry of Economy, Industry, and Competitiveness (MEIC) for the EMBL partnership, and grants 'Centro de Excelencia Severo Ochoa 2013–2017' SEV-2012-0208, and BFU2015-67107 co-founded by European Regional Development Fund (ERDF); from the CERCA Program/Generalitat de Catalunya; from the Catalan Research Agency (AGAUR) SGR857, and grants from the European Union's Horizon 2020 Research and Innovation Program under the Grant Agreement No. ERC-2016-724173. TG also receives support from a INB grant (PT17/0009/0023 – ISCIII-SGEFI/ERDF). CL-F received funding from Christian Doppler Laboratory for Fungal Infections: Avoid, find, and treat!

6.10 Acknowledgements

The authors thank all Gabaldón lab members for the helpful discussions and comments on this work, specially Marina Marcet-Houben for the help in the phylome analysis.

6.11 Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00383/full#supplementary- material

7 Whole-genome analysis reveals that hybridization is widespread among *Candida* species

Mixão, V., Boekhout, T., & Gabaldón, T. Whole-genome analysis reveals that hybridization is widespread among Candida species. (In preparation)

7.1 Abstract

Candida species are among the most important fungal pathogens. Recently, hybridization has been suggested to play a role on the emergence of some of them. Specifically, a hybrid nature has been described for most of all strains of *Candida orthopsilosis*, *Candida metapsilosis* (members of the *Candida parapsilosis* species complex), and *Candida inconspicua* (member of the *Pichia cactophila* species complex). To gain a better understanding of the origin of these hybrid pathogens, we sequenced the genome of eight close relatives of these two clades. Our results revealed two additional hybrid lineages in the CUG-Ser clade: *Candida theae* and *Candida subhashii.* Furthermore, we found that contrary to what was previously suggested, *P. cactophila* and *C. inconspicua* are not the same species. Instead our genomic analysis indicates that *P. cactophila* is a hybrid which has *C. inconspicua* as putative parental. Overall, these results point to a high propensity of *Candida* spp. to hybridize.

Keywords: Hybridization, *Candida*, *Pichia*, pathogens

7.2 Introduction

Candida species are the most important pathogenic yeasts and have been responsible for several outbreaks (Pfaller et al., 2010b; Turner & Butler, 2014). *Candida* infections (i.e. candidiasis) are mostly

caused by *Candida albicans, Candida glabrata,* and *Candida parapsilosis*, which combined account for more than 90% of the cases (Turner & Butler, 2014). However, in the last years the epidemiology of this disease has shifted with the emergence of new pathogens (Pfaller et al., 2010b). The mechanisms involved in the emergence of pathogenicity are still unknown, but hybridization has been suggested as a possible evolutionary path for the appearance of novel yeast pathogens (Mixão & Gabaldón, 2018). Indeed, in the *C. parapsilosis* species complex at least five natural independent hybridization events have been described, all of them comprising clinical isolates (Pryszcz et al., 2014, 2015; Schröder et al., 2016). Four of these events correspond to *Candida orthopsilosis* and the remaining one to *Candida metapsilosis* (Pryszcz et al., 2015; Schröder et al., 2016). More recently, the hybrid nature of the emerging pathogen *Candida inconspicua,* member of the *Pichia cactophila* species complex, has also been reported (Mixão et al., 2019). While one of the parental lineages for *C. orthopsilosis* hybrids has already been isolated, also in clinical environment, for *C. metapsilosis* and *C. inconspicua* none of the parentals has been found (Mixão et al., 2019; Pryszcz et al., 2015; Schröder et al., 2016). This has led to the hypothesis that these parentals are possibly non-pathogenic and that the hybridization event contributed to the emergence of pathogenic lineages (Mixão et al., 2019; Pryszcz et al., 2015). Therefore, it is important to analyze the genomes of closely related species in order to understand the evolution of these hybrids and, if possible, identify their putative parentals.

Candida jiufengensis, *Candida pseudojiufengensis*, *Candida oxycetoniae*, *Candida theae,* and *Candida subhashii* are five recently described species isolated from different environments, that belong to the CUG-Ser clade (Adam et al., 2009; Chang et al., 2012; Ji et al., 2009). While *C. jiufengensis, C. pseudojiufengensis* and *C. oxycetoniae* were isolated from the gut of flower beetles in China (Ji et al., 2009), *C. theae* was described in 2012 from isolates obtained from modern tea drinks and ancient chicha fermentation vessels in Taiwan and Ecuador, respectively (Chang et al., 2012). *Candida subhashii* was described in 2009 in Canada as a novel species causing peritonitis in humans and reported as a non-fermentative yeast very closely related to *C. parapsilosis* and *Candida tropicalis* (Adam et al., 2009). Later on, it was isolated from agricultural soil in Switzerland (Hilber-Bodmer et al., 2017). Although its first

description was as a fungal pathogen (Adam et al., 2009), it was suggested that the soil can possibly be its natural habitat, specially taking into account that it is a highly competitive yeast, behaving as antagonist for several filamentous fungi (Hilber-Bodmer et al., 2017). In the same work, the authors described a highly conserved mitochondrial genome between Canadian and European isolates from this species, suggesting a recent and fast global spreading. The mitochondrial genome of *C. subhashii* is particularly interesting since it is enriched in GC nucleotides, turning the species in the known yeast with the highest GC content on its mitochondrial DNA (Fricova et al., 2010). Given the putative proximity of these five species to the hybrid lineages of *C. orthopsilosis* and *C. metapsilosis*, we hypothesized that among them there could be a putative parental of these hybrids, or, at least, they could help to understand their evolution. Therefore, their genomes were sequenced and analyzed for the first time.

As mentioned above, *C. inconspicua* was recently described as a hybrid species (Mixão et al., 2019). The analysis of the genome of multiple clinical isolates from Europe revealed the existence of two clades which differ in terms of genomic variability and ploidy levels. Despite this difference between the two clades, it was not clear whether they corresponded to the same or different hybridization events (Mixão et al., 2019). *C. inconspicua* belongs to the *P. cactophila* species complex, which also includes *P. cactophila* and *Pichia norvegensis* (syn. *Candida norvegensis*). While *C. inconspicua* and *P. norvegensis* are considered emerging pathogens, *P. cactophila* has never been associated to human infections, being more frequently found in cacti (Guitard et al., 2013; Moraes et al., 2005; Sandven et al., 1997). However, a recent study suggested that *C. inconspicua* and *P. cactophila* correspond to the same species, even if they displayed differences in their ability to sporulate, which is almost nonexistent in the case of *C. inconspicua* (Guitard et al., 2015). To get a better insight on the evolution of this emerging pathogen, and clarify whether *P. cactophila* could represent one of the two previously described clades of this species, we here sequenced for the first time the type strain of *P. cactophila* (CBS6926), and performed a comparative genomics analysis between these two species and *P. norvegensis*. Moreover, given that the previous studies on *C. inconspicua* only included isolates from

Europe, an additional putative *C. inconspicua* isolate from Canada was also analyzed.

7.3 Results

The genomic variability in five new species of the CUG-Ser clade

For the analysis of the five species of the CUG-Ser clade, different short-read sequencing libraries were obtained (see Material and Methods). After applying different assembly strategies, the best assembly for each species was chosen based on different quality parameters such as genome completeness, percentage of mapped reads, N50, and fragmentation. The different metrics of the best genome assembly for each of the species are summarized in Table 7_1, which also includes information on the assembly strategy. Final *C. jiufegensis*, *C. pseudojiufegensis* and *C. oxycetoniae* assemblies have 13.83 Mb, 15.98 Mb, and 11.31 Mb, respectively. This high variability in the genome size of these three species was unexpected, given their close relatedness (Ji et al., 2009). Nevertheless, our estimations report > 99% of assembly completeness for each of them. The expected and observed genome sizes of *C. pseudojiufegensis* were higher than observed for any of the other two related species. However, a *k*-mer comparison revealed that the assembly does not have duplicated regions (Figure S1). Furthermore, the alignment of the assembly to itself discarded the presence of multiple haplotypes for the same region (Figure S2). Therefore, the higher genome size observed for *C. pseudojiufegensis* is unlikely the result of assembly artifacts. The reduced genome size of *C. oxycetoniae* is also of note. For this species, although the assembly completeness and the *k*-mers comparison did not reveal any significant missing portion of the genome (Table 7_1 and Figure S1), the predicted proteome completeness and the number of mapped reads were 96% and 97%, respectively (Table 7_1), suggesting that the assembly misses a small fraction of the genome. This conclusion is reinforced by the difference between the estimated and obtained genome sizes (12.82 and 11.31 Mb, respectively). Nevertheless, no other assembly strategy provided better results, and therefore we used this assembly for further analysis.

Table 7_1. Summary of the assembly metrics for the analyzed species of the CUG-Ser clade, namely, *C. jiufegensis, C. pseudojiufegensis, C. oxycetoniae, C. theae* and *C. subhashii.* Information on expected and observed genome sizes, number of contigs, N50, GC content, assembly completeness, number of predicted protein-coding genes, mapped reads, genomic variability, and assembly strategy are provided.

	C. jiufegensis	***C. pseudojiufegensis***	***C. oxycetoniae***	***C. theae***	***C. subhashii***
Estimated size (Mb)	13.83	15.98	11.31	12.36	15.61
Assembly size (Mb)	13.70	14.03	12.82	12.45	15.43
Contigs	22	56	194	181	345
Contigs >50kb	15	31	85	141	252
N50	1,273,379	828,659	149,712	205,21	108,455
GC	27.22 %	28.22 %	37.76 %	40.23 %	34.56 %
Assembly completeness (KAT)	99.94%	99.68%	99.83%	99.25%	58.81%
Proteome completeness (BUSCO)	98.90%	99.10%	96.00%	97.50%	99.30%
Protein number	5,673	5,814	4,92	5,407	6,178
Mapped reads	99.88%	99.43%	96.99%	99.35%	98.63%
SNPs/kb (heterozygous)	0.03 (0.02)	0.09 (0.09)	0.07 (0.04)	6.36 (6.34)	14.26 (14.23)
Parental divergence	-	-	-	3.15%	4.32%
LOH	-	-	-	84.41%	70.02%
Assembly strategy	SOAPdenovo + Redundans	dipSPAdes + Redundan	SOAPdenovo + Redundans	SPAdes + Redundans	SPAdes + Redundans

Contrary to *C. jiufengensis, C. pseudojiufegensis* and *C. oxycetoniae*, the *k*-mer counts in sequencing data of both *C. theae* and *C. subhashii* revealed the presence of at least two peaks of coverage (Figure 7_1A), which was indicative of the presence of highly heterozygous regions in their genomes, possibly resulting in highly fragmented genome assemblies. Indeed, our estimations suggest that both species are diploid (nQuire histotest $r^2 = 0.97$ for *C. theae* and $r^2 = 0.92$ for *C. subhashii*). For *C. theae*, a combination of paired-end and mate-

pair sequencing libraries generated a 12.45 Mb genome assembly with a completeness of 99.25%. The genome assembly of *C. subhashii* comprises 15.43 Mb, which corresponds to an assembly completeness of 58.81% (Table 7_1), a value expected for reduced genome assemblies of highly heterozygous species (Mixão et al. 2019; Pryszcz et al. 2014, 2015). It is worth noting that indeed for both species the assembly reduction was successful with approximately half of the *k*-mers of heterozygous regions being absent from the assemblies (Figure 7_1A), suggesting that only one of the haplotypes was kept. However, while this assembly reduction did not present any impact in the estimated assembly completeness of *C. theae*, for *C. subhashii* the estimated assembly completeness was low. As the estimated predicted proteome completeness of *C. subhashii* was > 99%, we conclude that the low assembly completeness is indeed related to the haplotype reduction, and not to a missing portion of the genome. This difference in the different estimations between the two species may suggest that i) *C. theae* is less heterozygous than *C. subhashii* (confirmed by the proportion of the two *k*-mer peaks in each strain, see Figure 7_1A), and ii) that the sequence divergence of the two haplotypes is higher in *C. subhashii*.

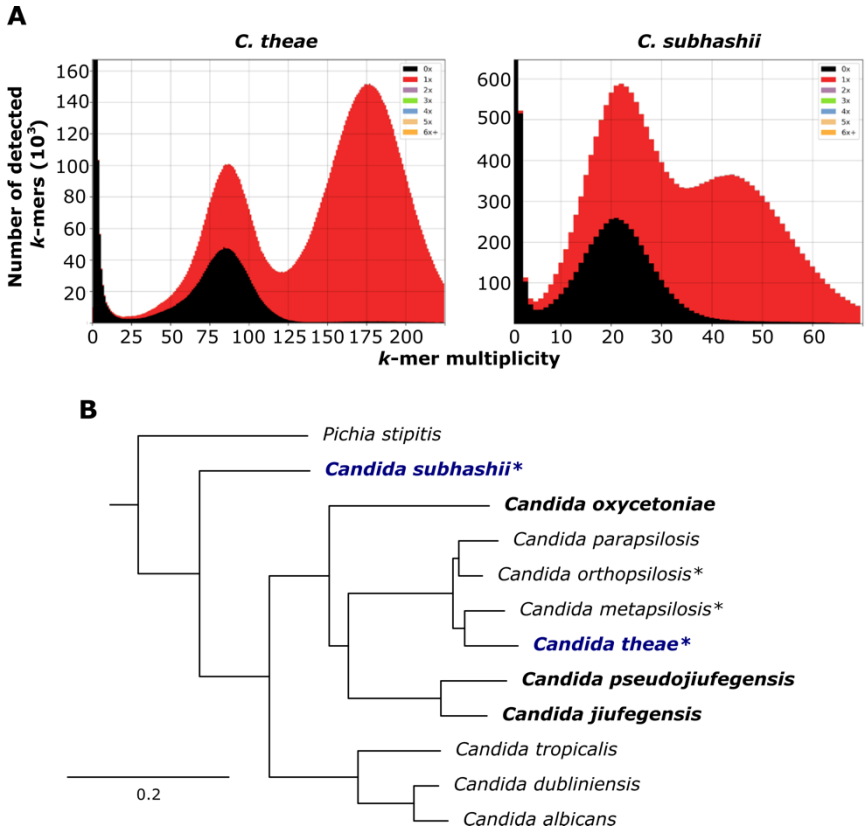

Figure 7_1. Analysis of the genome of the two new hybrid species of the CUG-Ser clade. **A)** *k*-mer comparison of the sequencing libraries of *C. theae* (on the left) and *C. subhashii* (on the right) with the respective genome assemblies. The *x*-axis represents the read coverage and the *y*-axis the *k*-mer frequency. *k*-mers that are present in single copy in the genome assembly are marked in red, while *k*-mers absent from the genome assembly are in black. **B)** Phylogenetic tree with the five species of the CUG-Ser clade analyzed in this work highlighted in bold. Hybrid lineages are marked with an asterisk, and hybrids described in this work for the first time are highlighted in blue.

The assessment of genomic variability was based on polymorphic positions. *C. jiufegensis*, *C. pseudojiufegensis* and *C. oxycetoniae* have less than 0.1 SNPs/kb, thus being highly homozygous species (Table 7_1). Consistent with our observations from the genome assemblies, the genomic variability of *C. theae* and *C. subhashii* is much higher. *C. theae* presents 6.34 heterozygous SNPs/kb, and *C. subhashii* 14.23 heterozygous SNPs/kb (Table 7_1). These results, indicating a much higher heterozygosity for *C. subhashii* as

compared to *C. theae*, are consistent with the observations from the assemblies discussed above. Importantly, for both species, these heterozygous variants are not homogeneously distributed along the genome, but rather they form blocks of high heterozygosity separated by homozygous regions. The estimated sequence divergence in the heterozygous blocks has a normal distribution in both cases (Figure S3). This resembles the patterns previously described for *Candida* hybrids (Mixão et al., 2019; Pryszcz et al., 2015; Schröder et al., 2016), and suggests that these heterozygous blocks are the footprints of past hybridization events, and not the result of accumulation of mutations in the same lineages across time. The estimated current sequence divergence between the two haplotypes is 3.15% and 4.32% for *C. theae* and *C. subhashii*, respectively, which is close to the estimated sequence divergence of the parentals of *C. metapsilosis* and *C. orthopsilosis* (approximately 4.5%, (Pryszcz et al., 2015; Schröder et al., 2016), Table 7_1). The homozygous regions are possible tracks of loss of heterozygosity (LOH), and both species present high levels of LOH, with 84.41% and 70.02% of *C. theae* and *C. subhashii* genomes, respectively, being homozygous (Table 7_1). It is important to note, that a *k*-mer comparison between the five species here analyzed and the previous hybrids of the *C. parapsilosis* species complex (*C. metapsilosis* and *C. orthopsilosis*), did not reveal any shared *k*-mer, suggesting that no species represent a possible parental lineage of any of the hybrid species. Moreover, a phylogenetic analysis revealed that *C. subhashii* is not as close to the *C. parapsilosis* species complex as previously thought (Figure 7_1B).

The genomic variability in the *P. cactophila* species complex

For a better understanding of the evolution of the hybrid pathogen *C. inconspicua*, we sequenced the type strain of *P. cactophila* (CBS 6926), and the genome of a putative *C. inconspicua* isolate from Canada, which is the first sequenced isolate outside Europe and is hereafter referred as the Canada strain. Contrary to what was performed for the species of the CUG-Ser clade, given the expected similarity between these two isolates and the type strain of *C. inconspicua*, we first compared and aligned the reads of these strains to the available assembly (Mixão et al., 2019). The *k*-mer comparison analysis revealed that the Canada strain has three peaks of coverage

(Figure 7_2A), suggesting that it is highly heterozygous, as expected, but also presents a ploidy different from 2. Indeed, our estimations suggest that this strain is triploid (nQuire histotest r^2 = 0.98). An important observation was the absence of the majority of the *k*-mers of this library in the genome assembly of *C. inconspicua*, suggesting that they do not represent exactly the same lineage. Indeed, this strain has 36.39 variants/kb when compared to the *C. inconspicua* reference genome, from which 29.04 correspond to heterozygous positions. These levels of heterozygosity are surprising when compared to the other *C. inconspicua* isolates analyzed in the past (minimum 14, and maximum 19.76 heterozygous variants/kb (Mixão et al., 2019)), but indicate that the Canada strain is also a hybrid. Furthermore, 37.94% of its genome corresponds to LOH regions. Once again, this value is compatible with a higher level of heterozygosity in this strain when compared to the previously analyzed isolates of *C. inconspicua* (minimum 51% LOH (Mixão et al., 2019)). The current nucleotide divergence between the homeologous chromosomes of this strain is 4.46%, ~1% higher than the *C. inconspicua* strains. These differences in the levels of LOH and parental sequence divergence are unexpected for a *C. inconspicua* isolate, but in accordance with the observations of the *k*-mer analysis. These results suggest that the Canada strain descends from a hybridization event unrelated to any of the previously described clades of *C. inconspicua*.

Similarly to the analysis of the Canada strain, *P. cactophila* was analyzed based on a read mapping approach on *C. inconspicua* genome. The *k*-mer comparison between these species revealed that indeed part of *P. cactophila* is represented in *C. inconspicua* genome assembly, but the most part of it is not (Figure 7_2A). We noted the occurrence of a single peak in the *k*-mer plot, which can be indicative of a homozygous species, or a highly heterozygous lineage with very low levels of LOH. Our estimations suggest that *P. cactophila* is diploid (nQuire histotest r^2 = 0.98), thus the single peak must correspond to a highly heterozygous lineage with low levels of homozygosity. Indeed, this species has 47.52 variants/kb when compared to *C. inconspicua* genome assembly, from which 39.52 are heterozygous. Interestingly, once again the heterozygous SNPs of this lineage are not homogeneously spread around the genome, instead forming blocks of heterozygosity with an estimated current nucleotide sequence divergence of 6.41%, ~3% higher than *C. inconspicua* isolates (Mixão et al., 2019), and ~2% higher than the

Canada strain. This supports a scenario of a different hybridization event. Of note, *P. cactophila* has 40.84% of the genome in LOH regions. The higher sequence divergence observed in *P. cactophila* indicates that the two parental lineages of this species are not exactly the same as *C. inconspicua* nor as the Canada strain, but the observation of shared *k*-mers between the two species suggests that they share at least one parental lineage. As can be seen in the phylogenetic tree shown in Figure 7_2B, the three putative *C. inconspicua* clades (including the Canada strain) form different hybridization clades, and *P. cactophila* is distantly related to any of them. These results contradict the previously proposed scenario that *P. cactophila* and *C. inconspicua* correspond to the same species (Guitard et al., 2015).

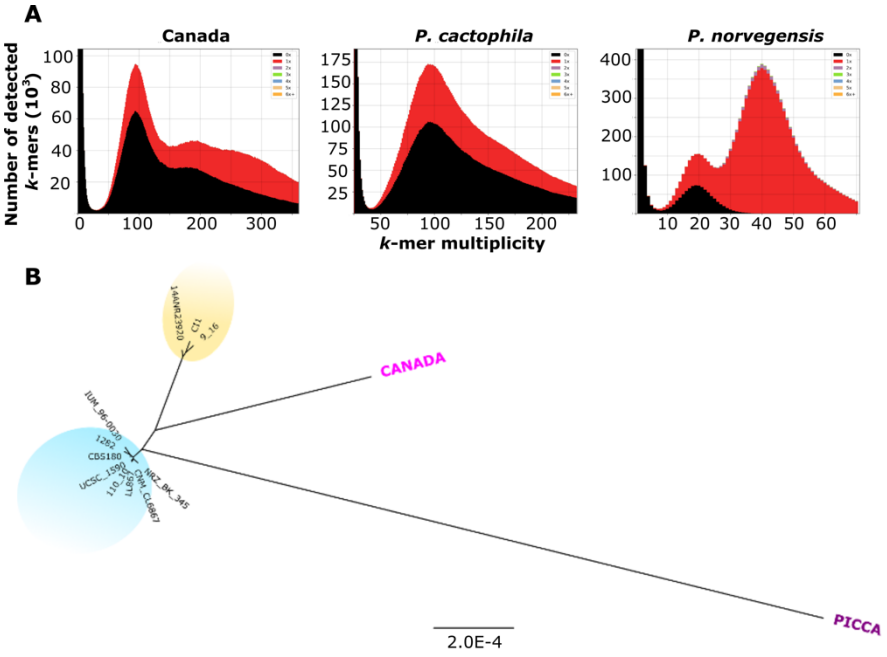

Figure 7_2. Analysis of the genome of the three new hybrid lineages of the *P. cactophila* species complex. **A)** *k*-mer comparison of the sequencing libraries of the Canada strain (on the left), *P. cactophila* (in the middle) and *P. norvegensis* (on the right) with the reference genome of *C. inconspicua*. The *x*-axis represents the read coverage and the *y*-axis the *k*-mer frequency. *k*-mers that are present in single copy in the genome assembly are marked in red, while *k*-mers absent from the genome assembly are in black. **B)** Phylogenetic tree with the four main lineages related to *C. inconspicua*. Blue and yellow highlight the previous described clades of *C. inconspicua* (Mixão et al., 2019). Pink represents the Canada strain, and purple *P. cactophila* (PICCA).

Analysis of four marker genes and the *MAT* locus

To get a better understanding of the number of lineages involved in the origin of the new hybrids here described, further analyses on four universal marker genes (*KOG1*, *CLU1*, *VPS53* and *RFA1*) (Capella-Gutierrez et al., 2014) and on the *MAT* locus were performed. Briefly, all these genes were phased in order to recover the two haplotypes in each heterozygous isolate and then reconstruct their phylogenetic relationships (see Material and Methods). Of note, in the case of the universal marker genes, due to the absence of known parental lineages, it was not possible to concatenate the different genetic information, and the analysis was performed separately for each gene. In the case of *C. theae*, and its close relatives *C. metapsilosis* and *C. orthopsilosis*, the reconstruction of the current haplotypes of the four marker genes confirmed that they do not share any parental lineage, and therefore this new hybridization event uncovers two additional unknown lineages (*C. theae* parentals) in the *C. parapsilosis* species complex (Figure 7_3A and Figure S4). Regarding *C. subhashii*, as mentioned before, it is not as related to any of the analyzed lineages as initially thought, therefore due to the absence of closely related species to this lineage the analysis of the phased four marker genes was not performed. As expected, for both species the analysis of the *MAT* locus revealed that they have *MAT* **a** and α alleles, consistently with their hybrid nature.

Regarding the *P. cactophila* species complex, the reconstruction of the different haplotypes in the marker genes revealed four putative clades (Figure S5). Two of them correspond to *C. inconspicua* parentals, which are also one of the parentals of the Canada strain and *P. cactophila*, and the remaining two correspond to the respective alternative parentals of these last two lineages. In the previous study in which multiple *C. inconspicua* isolates were screened (Mixão et al., 2019), it was suggested that the polymorphisms between the two *C. inconspicua* clades observed in the *MAT* locus could be a consequence of accumulation of mutations after a shared hybridization event, and not differences in their parental lineages. These results did not allow the exclusion of the hypothesis that the two clades possibly correspond to diverged lineages of the same hybridization event (Mixão et al., 2019). However, the current analyses show that the Canada strain is a result of an independent hybridization event, and even so we observe that it shares many of

the SNPs present in *C. inconspicua* clade 1 *MAT* **α**. Thus, these SNPs were possibly present before both hybridization events (the Canada strain and *C. inconspicua* clade 1), and therefore *C. inconspicua* clades 1 and 2 represent two different hybridizations. Furthermore, if the majority of these SNPs are ancestral, these results suggest that the Canada strain and *C. inconspicua* clade 1 share the "*MAT* **α** parent".

As mentioned above, the reconstruction of the different haplotypes in the set of marker genes revealed that one of the haplotypes of *P. cactophila* is close to *C. inconspicua*, but the alternative one is distantly related (Figure S5). This result supports the hypothesis that *P. cactophila* and *C. inconspicua* share a parental lineage but differ in the other one. However, further inspection of the *MAT* locus revealed a much more complex scenario. Indeed, *P. cactophila MAT* **a** seems close to that of *C. inconspicua* clade 1. However, the *MAT* **α** allele has a *MAT* **α1** that is different from any *C. inconspicua* strain, and a *MAT* **α2** that is actually heterozygous. To confirm this observation, i.e that *P. cactophila* has two *MAT* **α** alleles, we assembled the genome of the *P. cactophila* strain (see Material and Methods). From this genome assembly, we were able to recover the full *MAT* **a** and *MAT* **α** alleles of *P. cactophila*. Our results show that one of the haplotypes of *MAT* **α2** is recombined in the *MAT* **a** and corresponds to *MAT* **α2** of *C. inconspicua* (Figure 7_3B). The existence of this recombination event supports a scenario in which one of the parentals of *P. cactophila* was a hybrid *C. inconspicua* that underwent a recombination event in the *MAT* locus. It is important to note that our estimations point to a diploid state of *P. cactophila*, which is not compatible with the idea of a hybrid parental. Therefore, we suggest that after the recombination event in the *MAT* locus, the *C. inconspicua* hybrid that gave origin to *P. cactophila* experienced a ploidy reduction before the hybridization event (Figure 7_3C).

Analysis of the genome of *P. norvegensis*

With all the lineages analyzed so far in the *P. cactophila* species complex being hybrids, we decided to extend the analysis to the other pathogenic member of the clade, *P. norvegensis,* in order to assess whether it could be the alternative parental of *P. cactophila.* For that, we retrieved the publicly available sequencing library and genome

assembly for this species. Similarly to the other members of the clade, the *k*-mer analysis of this species revealed the presence of two peaks of coverage (Figure 7_2A), a pattern that as mentioned before is a good indicator of a highly heterozygous genome. The genomic variability of this species revealed 4.03 heterozygous variants/kb, thus lower than the variability of *C. inconspicua* and *P. cactophila*. Interestingly, similarly to the other two species, these variants are not spread across the genome, but rather form blocks of heterozygosity (Figure S6). The sequence divergence in the heterozygous blocks of *P. norvegensis* also has a single peak (Figure S6). This indicates that the heterozygosity of this species was possibly acquired at a single time point, meaning that it may have a hybrid nature. The current sequence divergence of this species is 3%, and the respective levels of LOH are 91%. This indicates that *P. norvegensis* is much more homozygous than any other hybrid of the clade. Hence, *P. norvegensis* comprises another hybrid in the *P. cactophila* clade, but with high levels of LOH, which can point to an older hybridization event or to a massive LOH. Importantly, the analysis of the *k*-mer comparison revealed that *P. norvegensis* is not the parental of any of the other two species, but rather another hybrid lineage in the *P. cactophila* clade which is emerging as a pathogen.

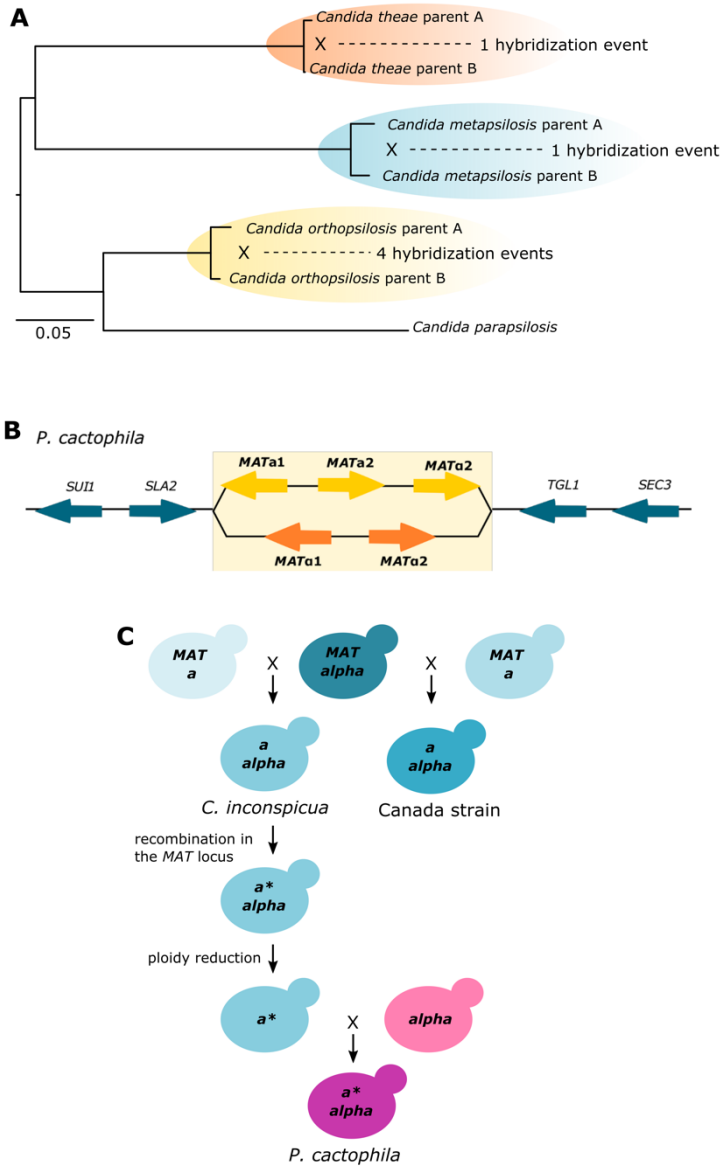

Figure 7_3. Summary of the different scenarios of hybridization described in this study. **A)** Phylogenetic tree of the different haplotypes of hybrid lineages of the *C. parapsilosis* species complex for *VPS53*. **B)** Schematic representation of the *MAT* locus of *P. cactophila*, in which the yellow genes belong to *C. inconspicua* hybrid parental (with a recombination) and the orange ones to the alternative parental lineage. **C)** Scheme of the different hybridization events involving *C. inconspicua* and *P. cactophila*. * represents the recombined **a** allele.

7.4 Discussion

In the last years, the emergence of new pathogenic lineages has shifted the epidemiology of candidiasis (Pfaller et al., 2010b). Species which so far were not considered as medically relevant are now emerging as new pathogens. For this reason, it is important to study those species and understand their evolution. Recent studies have pointed to a role of hybridization on the emergence of new pathogens, with different hybrid lineages being described in clinical isolates (Mixão et al., 2019; Mixão & Gabaldón, 2018; Pryszcz et al., 2015). These comprise lineages of the *C. parapsilosis* species complex (members of the CUG-Ser clade), and also of the *P. cactophila* species complex (Mixão et al., 2019; Pryszcz et al., 2014, 2015; Schröder et al., 2016). The parental lineages of these hybrids are still unknown, leading to the hypothesis that they are non-pathogenic (Pryszcz et al., 2015). To clarify this question, one of the major goals of this study was to analyze closely related species to these hybrid *Candida* lineages in order to find possible parentals. Thus, we explored the genomes of eight lineages, namely, *C. jiufegensis, C. pseudojiufegensis, C. oxycetoniae, C. theae* and *C. subhashii* in the CUG-Ser clade, and *P. cactophila*, *P. norvegensis* and a putative *C. inconspicua* isolate from Canada in the *P. cactophila* species complex. Unfortunately, none of these species represents the parental of the previously described hybrids, and it remains still unclear what was the role of hybridization in the emergence of their pathogenicity.

The analysis of the genome of the five members of the CUG-Ser clade revealed the presence of two additional hybrid lineages: *C. subhashii* and *C. theae.* While a phylogenetic analysis has placed *C. subhashii* far from the previously described hybrids in the *C. parapsilosis* species complex, it has also placed *C. theae* very close to *C. orthopsilosis* and *C. metapsilosis* hybrids (Figure 7_1B). These are remarkable results because they point to the existence of two additional unknown homozygous lineages in the species complex, which once again crossed and gave origin to a new species. This indicates a high propensity of these lineages to hybridize. *C. theae* represents an environmental isolate. Nevertheless, the high genomic plasticity characteristic of hybrid genomes, and the fact that it is evolutionary related to pathogenic species, should raise some concern

about the ability this species to infect humans. For instance, *C. subhashii*, here described as a hybrid species, was able to cause peritonitis despite its natural habitat being the soil (Adam et al., 2009; Hilber-Bodmer et al., 2017). This shows that environmental hybrids may have the ability to infect humans. Future studies should sequence and analyze the genome of environmental isolates of *C. subhashii* to understand if hybridization played an important role on its pathogenicity.

In the case of the *P. cactophila* species complex, an even more intricate scenario was found, with at least four hybrid lineages occurring within the clade. Our results confirm that the previously described clades of *C. inconspicua* (Mixão et al., 2019) correspond to two different hybridization events. This resembles the case of *C. orthopsilosis* in which the multiple crosses of the same two lineages originated multiple pathogenic hybrid clades (Schröder et al., 2016). More interesting, one of the parental lineages of *C. inconspicua* clades, specifically the one providing *MAT* **α**, has crossed with an alternative lineage with a higher sequence divergence that that of the alternative parental of *C. inconspicua*, giving rise to the Canada strain. This result raises the question of the taxonomic identification of hybrid lineages. For instance, this clinical isolate from Canada was identified as *C. inconspicua*, but it does not result from the cross of the same two parental lineages. Therefore, it is uncertain whether they share the same phenotypic traits, and whether they should have the same taxonomic classification. This question may also be generalized to hybrids originated from the cross of the same lineages because the process of genome shaping between the two parental genomes may occur in different ways generating lineages with different genetic information, and different phenotypes.

The analysis of the *P. cactophila* type strain revealed that this species is the result of a hybridization event of a *C. inconspicua* hybrid and a yet unknown lineage. This shows that contrary to what was previously suggested (Guitard et al., 2015), *C. inconspicua* and *P. cactophila* are not the same species. The conclusion regarding the origin of *P. cactophila* was taken based on a recombination event in the *MAT* locus of this species, in which the *MAT* **α2** of *C. inconspicua* was recombined in the *MAT* **a** allele of the same species in the genome of *P. cactophila*. A similar event has been previously

described in *C. metapsilosis* hybrids (Pryszcz et al., 2015), thus suggesting that it may represent an important advantage for these lineages. A previous study has reported that the disruption of the *MAT* locus is important for hybrids to recover their fertility (Ortiz-Merino et al., 2017). Therefore, we hypothesize that this recombination event enabled a *C. inconspicua* hybrid to restore fertility and mate with *P. cactophila* alternative parental (Figure 7_3C). Given the diploid state of *P. cactophila* we also hypothesize that after the recombination in the *MAT* locus, the *C. inconspicua* parental of *P. cactophila* restored the haploid state. The contribution of such a recombination event in the *MAT* locus for the restoration of haploid state and fertility is still unknown. Thus, future studies should assess what is the relevance of this event for the emergence of new lineages.

As an attempt to determine the alternative parental lineage of *P. cactophila*, we analyzed the genome of *P. norvegensis.* This analysis revealed that this species is not the parental of *P. cactophila*, and actually it also represents a hybrid lineage. It is worth noting that this species has massive levels of LOH. LOH is usually taken as a proxy for the relative age of hybrid lineages (Schröder et al., 2016), thus *P. norvegensis* hybridization is likely much older than any of the other hybridization events of the *P. cactophila* species complex. As an alternative scenario, this species may have been subject to relevant stress, and suffered stress-induced LOH (Rosenberg, 2011). Altogether, these results show a high propensity of all these *Candida* species to hybridize. Given the genomic plasticity characteristic of hybrid genomes that allows them to adapt to new niches (Mixão & Gabaldón, 2018), this should raise some concern regarding the emergence of new hybrid pathogens, and its consequences for the epidemiology of candidiasis.

7.5 Material and Methods

Genomic DNA sequencing

Genomic DNA sequencing was performed for the type strain of *C. jiufegensis* (CBS 10846), *C. pseudojiufegensis* (CBS 10847), *C. oxycetoniae* (CBS 10844), *C. theae* (CBS 12239), *C. subhashii* (CBS

10753) and *P. cactophila* (CBS 6926), and for the Canadian isolate of *C. inconspicua.* A modified protocol from the MasterPure™ Yeast DNA Purification Kit was used to extract the DNA. In brief, samples were grown overnight in liquid YPD at 30˚C. Cells were pelleted and lysed with RNAse treatment at 65˚C for 15 min. After 5 min of cooling down on ice, samples were purified by the kit reagent by mixing, centrifugation and removal of the debris as described in the kit protocol. Further, samples were left at -20˚C with absolute ethanol for at least 2 h after which the DNA was precipitated for 30 min at 4˚C. The pellet was washed in 70% ethanol and left to dry. TE buffer was used to resuspend the DNA. Genomic DNA Clean & Concentrator kit was used for the final purification.

Whole-genome sequencing was performed at the Genomics Unit from CRG. Libraries were prepared using the NEBNext® DNA Library Prep Reagent Set for Illumina® kit (New England BioLabs) according to the manufacturer's protocol. Briefly, 1 µg of gDNA was fragmented by nebulization in Covaris to approximately 600 bp and subjected to end repair, addition of "A" bases to 3′ ends and ligation of Truseq adapters. All purification steps were performed using Qiagen PCR purification columns (Qiagen). Library size selection was done with 2% low-range agarose gels. Fragments with average insert size of 700 bp were cut from the gel, and DNA was extracted using QIAquick Gel extraction kit (Qiagen) and eluted in 30 µl EB. 10 µl of adapter-ligated size-selected DNA were used for library amplification by PCR using the Truseq Illumina primers. Final libraries were analyzed using Agilent DNA 1000 chip to estimate the quantity and check size distribution and were then quantified by qPCR using the KAPA Library Quantification Kit (KapaBiosystems, ref. KK4835) prior to amplification with Illumina's cBot. Libraries were loaded at a concentration of 2 pM onto the flow cell and were sequenced 2 x 125bp on Illumina's HiSeq 2500.

For *C. theae* a mate-pair library was also sequenced. For that, DNA was fragmented to sizes between 1 and 20 kb using a transposase that binds biotinylated adapters at the breaking point. Strand displacement was performed to "repair" the nicks left by the transposase. Fragment sizes of 3 to 6 kb were then selected on a 0.8% agarose gel and were then circularized. Non-circularized DNA was removed by digestion. The circular DNA was then mechanically sheared to fragments of 100 bp to 1 kb aprox. and the fragments

containing the biotinylated ends were pulled down using magnetic streptavidin beads and submitted to a standard library preparation. A final size selection on 2% agarose gel was done and fragments of 400 to 700 bp were selected for the final library. Final libraries were analyzed using Agilent High Sensitivity chip to estimate the quantity and check size distribution and were then quantified by qPCR using the KAPA Library Quantification Kit (ref. KK4835, KapaBiosystems) prior to amplification with Illumina's cBot. Libraries were sequenced 2 x 125 bp on Illumina's HiSeq 2500.

Public data

For the analysis of *P. norvegensis*, publicly available data was retrieved from the NCBI database. Specifically, we downloaded the genome assembly under the accession number ASM370546v1, and the sequencing library under the sequencing run SRR6476040. For *k*-mer comparisons to assess possible parental relationship, sequencing libraries of *C. orthopsilosis* (ERR380554), *C. metapsilosis* (ERR247393) and *C. inconspicua* (SRR8506592) were also downloaded.

Genome assembly

Next-Generation Sequencing data from all strains were inspected with FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Paired end reads were filtered for quality bellow 10 or size below 31 bp and for the presence of adapters with Trimmomatic v0.36 (Bolger et al., 2014). Mate-pair reads were filtered with NxTrim v0.4.1-53c2193 (O'Connell et al., 2015). After filtration, only reads clearly identified as paired end or mate pair by the respective programs were used for further work. The K-mer Analysis Toolkit (KAT, (Mapleson et al. 2017)) was used to count *k*-mer frequency and estimate the expected genome size. SOAPdenovo v2.04 (Luo et al., 2012) and SPAdes v3.9 in both SPAdes and dipSPAdes modes (Bankevich et al., 2012; Safonova et al., 2015) were used separately to perform the genome assembly. Afterwards, redundant contigs were removed from each assembly with Redundans v0.13c (Pryszcz & Gabaldón, 2016). The quality of the different assemblies was inspected with Quast v4.5

(Gurevich et al., 2013). Genome annotation was performed with Augustus v3.5 using *C. albicans* as model organism (Stanke & Morgenstern, 2005). The assembly completeness was estimated with KAT and BUSCO v3 (Mapleson et al., 2017; Waterhouse et al., 2019). The best assembly for each species was chosen based on the assembly completeness, genome size, N50 and number of scaffolds. Alignment of the different genome assemblies to themselves was performed with MUMmer v3.1 (Kurtz et al., 2004).

Read mapping and Variant calling

Read mapping was performed with BWA-MEM v0.7.15 (Li, 2013). Picard v2.1.1 (http://broadinstitute.github.io/picard/) was used to sort the resultant file by coordinate, as well as, to mark duplicates, create the index file, and obtain the mapping statistics. The mapping was inspected with IGV version 2.0.30 (Thorvaldsdóttir et al. 2013). Mapping coverage was determined with SAMtools v0.1.18 (Li et al. 2009).

Samtools v0.1.18 (Li et al. 2009) and Picard v2.1.1 (http://broadinstitute.github.io/picard/) were used to index the reference and create its dictionary, respectively, for posterior variant calling. GATK v3.6 (McKenna et al. 2010) was used to call variants with the tool HaplotypeCaller set with --genotyping_mode DISCOVERY -stand_emit_conf 10 -stand_call_conf 30 -ploidy 2 -nct 8. The tool VariantFiltration of the same program was used to filter the vcf files with the following parameters: --clusterSize 5 --clusterWindowSize 20 --genotypeFilterName "heterozygous" --genotypeFilterExpression "isHet == 1" --filterName "bad_quality" -filter "QD < 2.0 || MQ < 40 || FS > 60.0 || HaplotypeScore > 13.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0" --filterExpression "DP <= 20" --filterName "DepthofQuality". In order to determine the number of SNPs/kb, a file containing only SNPs was generated with the SelectVariants tool. Moreover, for this calculation only positions in the reference with 20 or more reads were considered for the genome size, and these were determined with bedtools genomecov v2.25.0 (Quinlan and Hall, 2010). Ploidy estimation was calculated with nQuire histotest (Weiß et al., 2018).

LOH block definition

To determine for each heterozygous strain the presence of LOH blocks, heterozygous and homozygous variants were separated. Then, the procedure applied and validated by Pryszcz et al. (2015). Briefly, bedtools merge v2.25.0 (Quinlan & Hall, 2010) with a window of 100 bp was used to define heterozygous regions, and by opposite, LOH blocks would be all non-heterozygous regions in the genome. Moreover, the minimum LOH size was established at 100 bp.

Analysis of four universal marker genes

For the reconstruction of the phylogenetic relationships between the different hybrid lineages, the four marker genes previously described by Capella-Gutierrez et al. (2014), namely, *KOG1*, *CLU1*, *VPS53* and *RFA1*, were phased in each hybrid strain using HapCUT2 (Edge et al., 2017). The different haplotypes of each gene were aligned with MAFFT v7 (Katoh & Standley, 2013) and trimmed with trimAL (Capella-Gutiérrez et al., 2009). RAxML v8 (Stamatakis, 2014) was used to reconstruct the Maximum Likelihood phylogenetic tree of each of the multi-sequence alignments, using the GTRCAT model.

Phylogenetic tree reconstruction

The phylogenetic tree reconstruction with RAxML v8 (Stamatakis, 2014) for *C. inconspicua, P. cactophila* and the Canada strain was performed using GTRCAT model and considering the concatenated alignment of all the homozygous positions in all the respective strains, based on the alignment of their genomic reads on *C. inconspicua* genome assembly (Mixão et al., 2019). The phylogenetic tree reconstruction with RAxML v8 (Stamatakis, 2014) for the members of the CUG-Ser clade was performed using the PROTGAMMALG model and considering the concatenated alignment of the amino-acid sequences of the four marker genes. Multiple-sequence alignment and trimming were performed as described in the previous section.

8 Genomic evidence for a hybrid origin of the yeast opportunistic pathogen *Candida albicans*

Mixão, V., & Gabaldón, T. (2020). Genomic evidence for a hybrid origin of the yeast opportunistic pathogen *Candida albicans*. *BMC Biology*, *18*, 48. doi: 10.1186/s12915-020-00776-6

8 Genomic evidence for a hybrid origin of the yeast opportunistic pathogen *Candida albicans*

8.1 Abstract

Background: Opportunistic yeast pathogens of the genus *Candida* are an important medical problem. *Candida albicans*, the most prevalent *Candida* species, is a natural commensal of humans that can adopt a pathogenic behavior. This species is highly heterozygous and cannot undergo meiosis, adopting instead a parasexual cycle that increases genetic variability and potentially leads to advantages under stress conditions. However, the origin of *C. albicans* heterozygosity is unknown, and we hypothesize that it could result from ancestral hybridization. We tested this idea by analyzing available genomes of *C. albicans* isolates and comparing them to those of hybrid and non-hybrid strains of other *Candida* species.

Results: Our results show compelling evidence that *C. albicans* is an evolved hybrid. The genomic patterns observed in *C. albicans* are similar to those of other hybrids such as *Candida orthopsilosis* MCO456 and *Candida inconspicua*, suggesting that it also descends from a hybrid of two divergent lineages. Our analysis indicates that most of the divergence between haplotypes in *C. albicans* heterozygous blocks was already present in a putative heterozygous ancestor, with an estimated 2.8% divergence between homeologous chromosomes. The levels and patterns of ancestral heterozygosity found cannot be fully explained under the paradigm of vertical evolution and are not consistent with continuous gene flux arising from lineage-specific events of admixture.

Conclusions: Although the inferred level of sequence divergence between the putative parental lineages (2.8%) is not clearly beyond current species boundaries in Saccharomycotina, we show here that all analyzed *C. albicans* strains derive from a single hybrid ancestor and diverged by extensive loss of heterozygosity. This finding has important implications for our understanding of *C. albicans*

evolution, including the loss of the sexual cycle, the origin of the association with humans, and the evolution of virulence traits.

Keywords: *Candida albicans*, Yeast, Pathogen, Hybrid, Genome

8.2 Background

Hybrids are chimeric organisms that originate from the cross between two diverged lineages (whether from the same or distinct species). At the time of hybridization, the divergence at the sequence level between the pairs of homeologous chromosomes is similar to the divergence between the parental genomes. Consequently, hybrid genomes have high levels of heterozygosity, which can subsequently be eroded through recombination-mediated conversion of homeologous sequences, resulting in loss of heterozygosity (LOH, (Mixão et al., 2019; Mixão & Gabaldón, 2018; Pryszcz et al., 2015; Samarasinghe et al., 2020)). Hybridization may produce organisms with unique phenotypic features, which may contribute to the successful adaptation to new niches, being often associated with speciation (Abbott et al., 2013; Dagilis et al., 2019; Gladieux et al., 2014; Mallet, 2007). Many hybrids have been described in animals and plants. Examples go from butterflies to birds, nematodes, or even sunflowers (Lunt et al., 2014; Mallet et al., 2007; Ottenburghs, 2019; Welch & Rieseberg, 2002b). In fungi, advances in next-generation sequencing have recently allowed the identification of many hybrid lineages as well (Morales & Dujon, 2012; Tusso et al., 2019), of which some have importance for biotechnology and food or beverage industries (Krogerus et al., 2018; Marcet-Houben & Gabaldón, 2015; Monerawela & Bond, 2018). Hybrids with medical relevance have also been described, particularly in *Cryptococcus* and *Candida* clades (Hagen et al., 2015; Marcet-Houben & Gabaldón, 2015; Mixão et al., 2019; Mixão & Gabaldón, 2018; Pryszcz et al., 2015; Samarasinghe et al., 2020), with earlier studies suggesting that hybridization might be related to the emergence of virulence traits in some of these species (Li et al., 2012; Mixão et al., 2019; Mixão & Gabaldón, 2018; Pryszcz et al., 2015).

Candida species are the most common causative agents of hospital-acquired fungal infections (Brown et al., 2012; Consortium

OPATHY & Gabaldón, 2019; Gabaldón & Carreté, 2016; Lass-Flörl, 2009; Pfaller & Diekema, 2007), accounting for 72.8 million opportunistic infections per year, with an overall mortality rate of 33.9% (Jordà-Marcos et al., 2007; Pfaller & Diekema, 2007). *Candida albicans* is a commensal organism that can form part of the microbiota of healthy individuals (Cauchie et al., 2017). Under certain circumstances, such as a weakening of the host immune system, *C. albicans* can shift from commensal to pathogenic behavior (Pfaller & Diekema, 2007). This species is the causative agent in more than 50% of the candidaemia cases worldwide (Pfaller & Diekema, 2007). Although *C. albicans* cannot undergo a normal sexual cycle involving meiosis, it is known to be able to follow a so-called parasexual cycle (Berman & Hadany, 2012; Forche et al., 2008). This consists of the mating of two diploid cells, forming a tetraploid cell that subsequently returns to a diploid state by concerted chromosomal loss. Restoration of the diploid state is not always properly achieved, leading to aneuploidies (Bennett, 2015; Berman & Hadany, 2012). Thus, this system constitutes a source of genetic variability, which has been proposed to be advantageous under stress conditions (Berman & Hadany, 2012). Although the ability to undergo a sexual or parasexual cycle has not been thoroughly investigated in non-albicans *Candida* species, accumulating evidence suggests that some forms of mating and genomic recombination might be common even in species traditionally considered as asexual (Alby & Bennett, 2010).

Recently, the genomic diversity of *C. albicans* strains belonging to different MLST-based clades and isolated from globally distributed locations was investigated (Bensasson et al., 2019; Ene et al., 2018; Hirakawa et al., 2015; Ropars et al., 2018; Wang et al., 2018). These studies have shown that the *C. albicans* genome shows signs of recombination, with genomic material exchanged between different strains (Ropars et al., 2018; Wang et al., 2018). Furthermore, they have reported that the fraction of the genome covered by heterozygous regions can vary between 48 and 89%, depending on the strain, and that these heterozygous tracts are separated by regions of LOH (Bensasson et al., 2019; Ene et al., 2018; Ropars et al., 2018; Wang et al., 2018). Moreover, it has been shown that the accumulation of mutations and the exchange of genomic material between strains are the main forces shaping *C. albicans* genome (Wang et al., 2018).

However, an intriguing and still unaddressed question is: what is the initial source of the high heterozygosity levels present in *C. albicans* genome? Can the accumulation of mutations over long periods of time and the presence of inter-strain recombination explain the levels of heterozygosity observed in highly heterozygous regions of *C. albicans* strains? We noted that the genomic patterns observed in *C. albicans* are reminiscent of recently analyzed yeast hybrid species, such as *Candida inconspicua*, *Candida orthopsilosis*, and *Candida metapsilosis* (Mixão et al., 2019; Pryszcz et al., 2014, 2015; Ropars et al., 2018; Schröder et al., 2016). Hence, a possible scenario for the observed genomic patterns in *C. albicans* is that the divergence observed in highly heterozygous regions is not exclusively the consequence of continuous accumulation of mutations within the lineage, but also, to a large degree, a footprint of an ancient hybridization event between two diverged lineages. We here put these alternative hypotheses at test by comparing *C. albicans* genomic patterns with the ones observed in *C. inconspicua*, *C. orthopsilosis*, and *C. metapsilosis* hybrid strains, as well as, non-hybrid strains from these and other species.

8.3 Results

K-mer profiles of *C. albicans* sequencing libraries reveal a heterogeneous content similar to that of hybrid genomes

To assess heterozygosity levels in *C. albicans* genomes, we analyzed 27-mer frequencies of raw sequencing data of Illumina paired-end libraries from different *C. albicans* strains (see the "Methods" section). All analyzed sequencing libraries produced similar 27-mer profiles, showing two peaks of depth of coverage, one with half coverage of the other, which corresponded to heterozygous and homozygous regions, respectively (Figure 8_1A and Additional file 1: Fig. S1). Of note, for all strains, including the reference, approximately half of the 27-mers of the first peak were not represented in the reference assembly (Figure 8_1A and Additional file 1: Fig. S1) and therefore could correspond to heterozygous regions where only one of the haplotypes was represented in the

reference assembly. This bimodal pattern was also produced by sequencing libraries from hybrid strains of *C. orthopsilosis* and *C. metapsilosis* mapped to their respective reference genomes (Figure 8_1A and Additional file 1: Fig. S1) and was previously reported for *C. inconspicua* hybrids (Mixão et al., 2019). However, this pattern was not observable in sequencing libraries from non-hybrid strains of *Candida dubliniensis*, *Candida tropicalis*, *C. orthopsilosis*, and *Candida parapsilosis* (Figure 8_1A and Additional file 1: Fig. S1). As shown in Additional file 1: Fig. S1, hybrid strains with lower levels of heterozygosity (i.e., strains from *C. orthopsilosis* clade 1) had a higher homozygous peak as compared to hybrid strains with higher levels of heterozygosity (i.e., *C. orthopsilosis* clade 4). Altogether, these results show that the relative frequency of the two peaks is representative of the level of heterozygosity in hybrid genomes and that the patterns observed in genomes from *C. albicans* strains are similar to those of hybrid strains of *C. orthopsilosis* clade 1 (e.g., strain MCO456), which underwent extensive levels of LOH after hybridization (Pryszcz et al., 2014; Schröder et al., 2016).

Heterozygosity patterns in *C. albicans* are comparable to those of hybrid lineages

We next assessed heterozygosity levels by aligning genomic reads of *C. albicans* strains, including putative wild strains isolated from oak trees (Bensasson et al., 2019), on the reference for haplotypes A and B, independently (see the "Methods" section). Similar results were obtained for the two haplotypes, and therefore, we only report the results for haplotype A. This analysis revealed a genome-wide average heterozygosity of 6.70 heterozygous variants per kilobase (kb), which depending on the clade varied from 4.59 to 8.62 variants/kb (Additional file 2: Table S1). Once again, these values are comparable to what was observed for *C. orthopsilosis* MCO456 (clade 1) strain, where we found 7.16 heterozygous variants/kb (Additional file 3: Table S2). Of note, *C. albicans* strains isolated from oak trees were highly heterozygous, as previously reported (Bensasson et al., 2019), but their values were in the range of what we observed for clinical isolates.

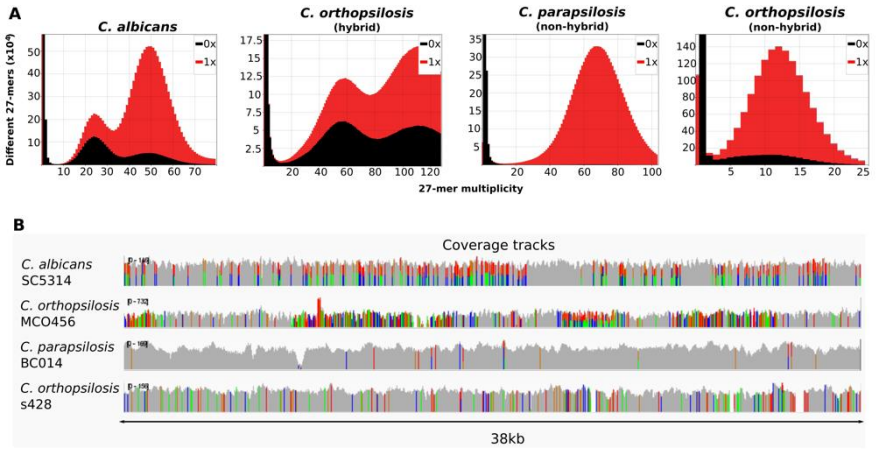

Figure 8_1. Comparison of the genomic patterns observed in *C. albicans* and hybrid and non-hybrid species. **A)** The 27-mer frequency for SC5314 (*C. albicans*), MCO456 (*C. orthopsilosis* hybrid), BC014 (*C. parapsilosis* non-hybrid), and s428 (*C. orthopsilosis*, non-hybrid parental lineage A), and their respective presence (red) or absence (black) in the respective reference genome (plots were obtained with KAT). **B)** Coverage tracks for illustrative genomic regions of the abovementioned genomic sequencing libraries, when aligned to the respective genomes. Colors indicate polymorphic positions. Positions with more than one color correspond to heterozygous variants. Visualizations were performed with IGV.

Furthermore, as previously described (Bensasson et al., 2019; Hirakawa et al., 2015; Ropars et al., 2018; Wang et al., 2018), the heterozygous variants in *C. albicans* were not homogeneously distributed across the genome. Rather, these variants were concentrated in heterozygous regions separated by what appeared to be blocks of LOH (Figure 8_1B and Additional file 4: Fig. S2). Moreover, we could identify some regions which were highly heterozygous in some strains whereas in others corresponded to LOH regions that tended to alternative haplotypes (Additional file 4: Fig. S2). These patterns were reminiscent of the ones observed in *Candida* hybrid species (Figure 8_1B and Additional file 4: Fig. S2) (Mixão et al., 2019; Pryszcz et al., 2014, 2015; Schröder et al., 2016).

The comparison of heterozygosity patterns between different species is often hampered by the use of different methodologies and criteria to define heterozygous and homozygous regions in different studies. For instance, while studies performed so far on *C. albicans*

distinguished heterozygous from LOH regions based on SNP density within windows of 5 kb (Ene et al., 2018; Hirakawa et al., 2015; Wang et al., 2018), 10 kb (Ropars et al., 2018), or even 100 kb length (Bensasson et al., 2019), studies performed on *Candida* spp. hybrids defined these blocks based on the distance between heterozygous positions (Mixão et al., 2019; Pryszcz et al., 2014, 2015). This last approach makes the boundaries between heterozygous and LOH blocks more flexible and precise and avoids averaging levels of heterozygosity when a window spans both homozygous and heterozygous regions. We therefore decided to use the methodological framework previously applied and validated in the study of hybrid species (see the "Methods" section) to analyze genome-wide heterozygosity patterns and define LOH blocks in *C. albicans* and compare them to patterns in established *Candida* hybrids.

On average, in *C. albicans* strains, we could define 7,059 heterozygous blocks and 16,492 LOH blocks, representing 11.18% and 85.74% of the genome, respectively. This is again notably similar to *C. orthopsilosis* hybrid clade 1, where 84.85% of the genome underwent LOH (Additional file 3: Table S2 and Additional file 5: Tables S3). Although these heterozygous blocks in *C. albicans* comprised most of the heterozygous SNPs (on average 59.35%), 12.79% of the heterozygous variants were placed within LOH blocks, and the remaining 27.86% in undefined regions (see the "Methods" section). The number of heterozygous variants outside heterozygous blocks was comparatively much higher than those found in *C. orthopsilosis* or *C. metapsilosis* hybrids (where it ranged from 2.38 to 5.81%, depending on the strain, Additional file 3: Table S2), but notably closer to that found in the recently identified *C. inconspicua* hybrids (up to 23.23%, (Mixão et al., 2019)). The higher number of heterozygous variants within LOH blocks would suggest that *C. albicans* and *C. inconspicua* LOH blocks have been accumulating mutations for a longer time, as compared to *C. orthopsilosis* or *C. metapsilosis*. In addition, it is worth noting that the union of the heterozygous blocks of the 61 *C. albicans* strains analyzed in this work corresponded to more than half (53.17%, Additional file 6: Table S4) of the *C. albicans* genome, and this value is expected to increase with a larger sample size. Although from these analyses, we cannot completely exclude the possibility that the accumulation of mutations followed by recombination between

different strains is responsible for the heterozygosity in *C. albicans* (Ropars et al., 2018; Wang et al., 2018), their strong similarity with what is observed for *C. inconspicua*, *C. orthopsilosis*, and *C. metapsilosis* hybrid strains and the high level of heterozygosity of *C. albicans* genome strongly suggest a scenario where hybridization between two diverged lineages was followed by extensive LOH.

The majority of heterozygous variants in *C. albicans* predate the diversification of known clades

The level of conservation of heterozygosity patterns across *C. albicans* strains of different clades can be used to assess the possibility of an ancestral hybridization event. Indeed, if a hybridization event between two divergent lineages, rather than the vertical accumulation of variants, was responsible for a sizable fraction of the heterozygosity levels observed across *C. albicans* genomes, then we would expect that a significant amount of heterozygous SNPs within heterozygous blocks would be shared by strains from deeply divergent clades, as their origin would have predated the diversification of the different *C. albicans* clades. To test this, we selected three non-overlapping sets of four strains (considered as replicates, see the "Methods" section) so that each set contains a representative strain of each of four deeply divergent clades (Figure 8_2A), according to the recent strain phylogeny described by Ropars et al. (2018). For each group, we compared the heterozygous positions in heterozygous and LOH regions shared by the different clades (Figure 8_2B, see the "Methods" section). We inferred that a heterozygous position in a given strain was ancestral if the most parsimonious reconstructed scenario (i.e., the one involving the lowest number of mutations) pointed to the same heterozygous genotype (i.e., the combination of the same two alleles) in the common ancestor of the four clades. The results, considering positions that could be unambiguously inferred, were consistent between the different groups (Figure 8_2C, Additional file 7: Table S5) and showed that on average, between 79.93 and 83.34% of the heterozygous positions within heterozygous blocks were ancestral (i.e., were present before the divergence of the clades), as compared to 15.28 to 20.10% of heterozygous positions within LOH blocks. Interestingly, the density of ancestral SNPs supporting a hybridization scenario presented a normal distribution

with a peak at 20 SNPs/kb in all strains (Additional file 8: Fig. S3). The high level of common SNPs in heterozygous regions is shared between different clades even when considering coding and non-coding regions separately (Additional file 7: Table S5). These results strongly suggest that most of the heterozygous variants in heterozygous blocks were already present in a putative *C. albicans* ancestor, whereas most of the variants in LOH blocks appeared later, by independent accumulation in the different lineages.

In agreement with these results, a maximum likelihood phylogeny of reconstructed haplotypes in heterozygous regions for the same groups of four strains (where A and B refer to SC5314 described haplotypes, see the "Methods" section) indicated that the phylogenetic distance between the two haplotypes is higher than the distance between the different strains (Figure 8_2D and Additional file 9: Fig. S4). A similar approach was used in the past to confirm a hybrid origin of *C. orthopsilosis* (Pryszcz et al., 2014). Of note, the specific topological arrangement between strains is different for each of the two haplotypes, which supports the existence of recombination between haplotypes.

Based on the number of variants per kilobase in heterozygous blocks, we estimated that the homeologous chromosomes are currently approximately 3.5% divergent at the nucleotide level (Figure 8_2E and Additional file 5: Table S3). This estimation was consistent across the sixty-one different strains (ranging from 3.44 to 3.59%, Additional file 5: Table S3) and therefore unrelated to their overall level of heterozygosity. These analyses strongly suggest that most of the divergence between haplotypes in *C. albicans* heterozygous blocks was already present in a putative, highly heterozygous ancestor, with an estimated 2.8% divergence between the homeologous chromosomes (assuming ~80% of the current variants in heterozygous blocks were heterozygous in the ancestor, in line with our estimations above). We consider that vertical accumulation of such level of heterozygosity across the entire genome is not plausible, as it would imply extremely long periods of mutation accumulation in the absence of any inter-strain recombination or LOH, events which have been shown to be common in this species (Ropars et al., 2018). Therefore, we consider that the most likely scenario to explain such a pattern is a

hybridization event between two divergent lineages thereby forming a highly heterozygous ancestor.

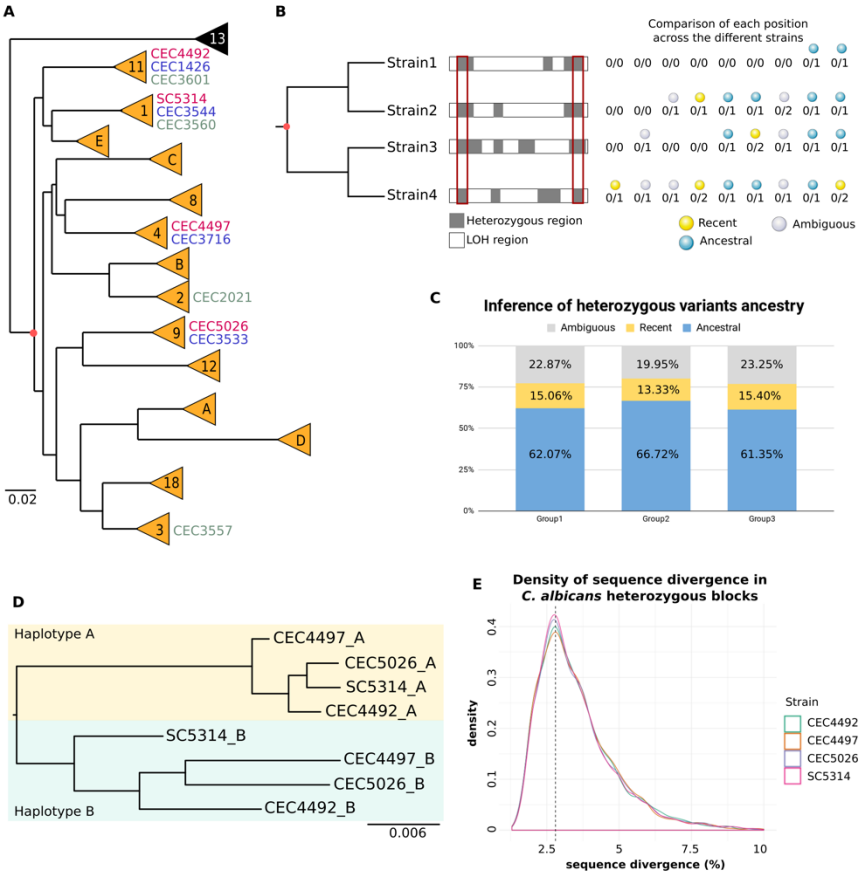

Figure 8_2. Analysis of *C. albicans* heterozygosity patterns. **A)** Schematic phylogenetic tree adapted from Gabaldón & Fairhead (2019) indicating the different *C. albicans* clades in orange. The potential common ancestor is marked with a red dot. Strains used for the comparison of heterozygous positions across different clades are indicated, with strains of the same group of analysis being written with the same color (black—group 1, blue—group 2, and gray—group 3). **B)** Schematic representation of the methodology for the comparison of heterozygous SNPs in heterozygous blocks (gray). The intersection of the heterozygous blocks is represented by the red rectangle (although not shown, the same approach was used for the analysis of LOH blocks). Examples of possible combinations of genotypes across the four strains are given ("0"—allele similar to the reference, "1"—allele different from the reference, "2"—allele different from the reference and from "1"). The most parsimonious path for the SNPs observed in each position was reconstructed. The decision taken for each position in a given strain is represented by yellow (recent), blue (ancestral), or gray (ambiguous)

spheres. **C)** Plot of the average proportion of ambiguously (gray) and unambiguously assigned positions for each group of strains. For unambiguously assigned positions, the proportion of recent and ancestral positions is shown in yellow and blue, respectively. **D)** Maximum likelihood phylogeny of the aligned reconstructed haplotypes A and B for the intersection of heterozygous blocks > 100 bp of group 1. **E)** Sequence divergence distribution in heterozygous blocks of *C. albicans* SC5314, CEC4492, CEC4497, and CEC5026.

Candida africana is not a putative parental of the hybrid ancestor of *C. albicans*

C. africana was proposed to be ranked as a species in 2001 (Tietz et al., 2001). However, although it presents particular phenotypes, the genetic similarity with *C. albicans* makes this controversial, and *C. africana* is often considered another *C. albicans* clade (clade 13, (Romeo et al., 2013)). In a recent population genomics study, Ropars et al., showed that contrary to *C. albicans*, *C. africana* is highly homozygous, and hypothesized that massive LOH might have occurred in this clade (Ropars et al., 2018).

Given our results indicating that *C. albicans* originates from a hybridization event, we wanted to investigate the possibility that *C. africana* lineage could correspond to one of the parentals involved in the hybridization. In the previously described *C. orthopsilosis* hybrids, for which one of the parental lineages is known, both *MATa* and *MATalpha* alleles in the hybrids exhibit a level of sequence divergence between the two parentals that is similar to that of the remaining nuclear genome (Sai et al., 2011; Schröder et al., 2016). Additionally, for each hybrid clade, when the two mating loci are present, only one of them is similar to the parental strain, with the other having high divergence indicating it descends from the other parental. Therefore, if *C. africana* was indeed one of *C. albicans* parentals, we would expect only one of the alleles to be inherited by the ancestral hybrid. In this case, we would expect only one of the two mating type loci in *C. albicans* clades to be similar to *C. africana*. Contrary to this expectation, our analysis reveals that both *MATa* and *MATalpha* of *C. africana* are highly similar (0.21% and 0.25% divergence, respectively) to those of *C. albicans*. This suggests that both *C. africana* and *C. albicans* share the same *MAT* locus alleles, and therefore, *C. africana* is not a parental species but rather another descendant from the same hybrid ancestor.

To confirm this, we selected a sample of *C. africana* strains (see the "Methods" section) and analyzed the respective genomic patterns. As expected, given the high levels of LOH previously described for this lineage (Ropars et al., 2018), *C. africana* presented low levels of heterozygosity, with an average of 2.68 heterozygous variants/kb, which were still high when compared to non-hybrid strains (Additional file 10: Table S6). Therefore, we decided to define heterozygous regions in *C. africana* strains. In contrast to *C. albicans*, only 3.8% of the genomes, on average, corresponded to heterozygous blocks. These blocks have a current haplotype divergence of 3.7%, which is close to the 3.5% mentioned above for *C. albicans* (Additional file 10: Table S6).

Furthermore, if *C. africana* was one of *C. albicans* parents, we would expect the homozygous regions of a given chromosome to tend to correspond always to the same *C. albicans* haplotype. The available phased genome of *C. albicans* is based on the heterozygous strain SC5314 (Muzzey et al., 2013), which already underwent LOH. Thus, only the regions of the phased reference genome corresponding to heterozygous regions of this strain can be used to assess distinct haplotypes in the inferred ancestral hybrid. From these regions, we selected heterozygous positions that were considered ancestral in the abovementioned analyses and compared them to homozygous regions in *C. africana* strains (see the "Methods" section for details). Our results indicate that similar proportions of homozygous positions in *C. africana* could be mapped to each of the two haplotypes (55% A and 45% B, where A and B refer to SC5314 haplotypes, Additional file 11: Table S7). This suggests that each *C. africana* chromosome is a mosaic of the two SC5314 haplotypes. Although, due to recombination between ancestral haplotypes (see above), SC5314 haplotypes might be chimeric in relation to the ancestral parental genomes. This result, together with the existence of heterozygous blocks in *C. africana* genome, provides support for a scenario considering massive LOH from a common highly heterozygous ancestor. A shared ancestral hybridization scenario is reinforced by the fact that for the majority (99%) of the ancestral *C. albicans* heterozygous positions, one of the alleles was represented in *C. africana.*

Continuous gene flux from divergent lineages cannot explain consistent patterns found across strains

Taken together, our results show compelling evidence for a highly heterozygous genome in the ancestor of currently sampled *C. albicans* and *C. africana* clades. The presence of highly divergent haplotypes and its consistency over the entire *C. albicans* genome can be best explained by an ancestral hybridization event between two distinct lineages and subsequent evolution through LOH. The alternative scenario of continuous introgression between divergent lineages could as well explain the presence of heterozygous regions in *C. albicans* strains but could not explain the observed similarity of heterozygosity patterns across strains (Figure 8_3). Indeed, in a single hybridization scenario followed by divergence, extant heterozygous blocks in divergent strains are expected to present similar levels of sequence divergence, to share the same ancestral heterozygous SNPs, and to show some degree of overlap in their patterns of heterozygosity, as we have described. Moreover, we expect similar levels of heterozygosity and LOH between the strains from the same hybridization event (Schröder et al., 2016). Such consistent features of heterozygous blocks across divergent strains are difficult to explain by the exclusive accumulation of lineage-specific and independent events of admixture, because sequence divergence is expected to be time dependent, and therefore, different events would leave different tracks (Figure 8_3C). Importantly, an ancestral hybridization scenario not only readily explains the heterozygosity patterns found in extant strains, but also could provide an explanation for the origin of other peculiar characteristics of *C. albicans* such as the absence of a standard sexual cycle, or its ubiquitous diploid nature, as discussed further below.

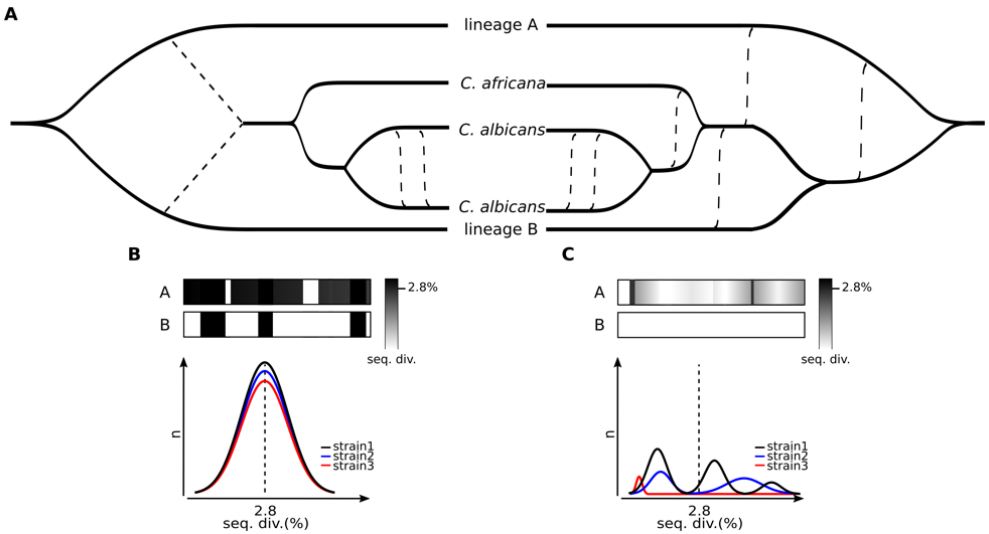

Figure 8_3. Schematic representation of plausible scenarios for the origin of *C. albicans* heterozygosity. **A)** The scenario proposed by this work is presented on the left, showing an ancestral hybridization event between two diverged lineages as the main source of *C. albicans* heterozygosity, followed by LOH, particularly extensive in *C. africana*, and more recent exchange of genomic material between *C. albicans* strains (dashed lines). The alternative scenario is presented on the right, showing inter-strain recombination (dashed lines) as the only explanation for the heterozygous patterns observed in *C. albicans*. **B)** Scheme of the expected sequence divergence patterns between the two haplotypes if a single hybridization event was in the origin of the heterozygosity in *C. albicans*. After a hybridization event, the heterozygous blocks are expected to have similar sequence divergence, which is then reflected in a normal distribution. This divergence is expected to be similar in all strains originated from the same event. **C)** Scheme of the expected sequence divergence patterns if inter-strain recombination was the only source of variability in *C. albicans*. In this scenario, the sequence divergence is time and strain dependent, and therefore, different patterns are expected between different blocks and between different strains

8.4 Discussion

Advances in next-generation sequencing have recently allowed the identification of many hybrid fungi with clinical relevance (Hagen et al., 2015; Mixão et al., 2019; Pryszcz et al., 2014, 2015; Schröder et al., 2016). Hybridization between diverged lineages is known to have an important role in the adaptation to new environments, or even in the emergence of new pathogens, as it is hypothesized to be the case of *C. metapsilosis* and *C. inconspicua* (Mixão et al., 2019; Mixão & Gabaldón, 2018; Pryszcz et al., 2015). The hybrid nature of these strains was discovered by noting the presence of highly heterozygous genomes with a large divergence between alternative haplotypes and showing characteristic non-homogeneous distributions of heterozygous variants, resulting in highly heterozygous blocks separated by regions of low heterozygosity, likely resulting from LOH events. In most such cases, these hybrid strains are diploid and seem to be unable to undergo a normal sexual cycle (Pryszcz et al., 2014).

C. albicans, the most important yeast pathogen for human health (Barnett, 2008), was previously shown to present genomic regions with high heterozygosity separated by what appeared to be blocks of LOH (Ropars et al., 2018; Wang et al., 2018). Parasex and admixture were taken as the possible source for the observed levels of heterozygosity (Ropars et al., 2018). However, as the description of these patterns was reminiscent of that observed in hybrid lineages (Mixão et al., 2019; Pryszcz et al., 2015; Schröder et al., 2016), we hypothesized that hybridization could have been the initial source of genomic variability in *C. albicans*. Our results show compelling evidence that *C. albicans* descend from a hybrid between two divergent lineages. The genomic patterns observed in this pathogen are similar to those of other hybrids, especially *C. orthopsilosis* MCO456 and *C. inconspicua* (Mixão et al., 2019; Pryszcz et al., 2014; Schröder et al., 2016). This scenario clearly points to a highly heterozygous ancestor predating the divergence of currently known clades. The reconstructed most recent common ancestor of sequenced *C. albicans* strains would present most (53.17%) of its genome within heterozygous blocks. Current heterozygous regions present a 3.5% divergence between the two haplotypes, and we here infer that around 80% of such variants are ancestral, suggesting that

the putative *C. albicans* ancestor had, at least, roughly 2.8% sequence divergence at the nucleotide level, and that this hybridization event is relatively older when compared to other hybrids, such as *C. orthopsilosis* or *C. metapsilosi*s (Pryszcz et al., 2015; Schröder et al., 2016). From our analyses, we conclude that the existence of such level of divergence between two haplotypes in heterozygous regions can be better explained by these two haplotypes being genetically isolated for a long time. A hybridization between two previously isolated lineages, followed by LOH and further accumulation of SNPs, would readily explain the observed patterns in *C. albicans*. Alternative scenarios accounting for the origin of heterozygous regions through independent introgression events would not explain the widespread presence of conserved heterozygous SNPs across strains from deeply divergent clades, nor the normal distribution observed for the levels of sequence divergence between haplotypes (Figure 8_2E).

We want to stress that the scenario of the hybridization between divergent lineages is agnostic to the consideration of the parental lineages as different or the same species. What is relevant is the realization of genomic chimerism predating the divergence of *C. albicans* clades. The species concept in microbes is controversial. The estimated 2.8% ancestral divergence between the hybridizing lineages largely exceeds levels of divergence found between most distantly related strains of well-studied yeast species such as *Sacharomyces cerevisiae*, where 1.1% sequence divergence was found between the most distantly related strains (Peter et al., 2018), and is higher than the estimated divergence between different described fungal species such as 1.4% between *Verticillium dahliae* and *Verticillium longisporum* D1 parental (Inderbitzin, et al., 2011a, 2011b). On the other hand, it can be considered low when compared to ~4.6% divergence between distant strains in *Saccharomyces paradoxus* (Liti et al., 2006). Independently of the consideration of this putative ancestor as an inter- or intra-species hybrid, our results indicate that the ancestral hybrid did not backcross significantly with any of the parental lineages but rather further evolved in a mostly clonal manner.

Inability to undergo meiosis and to complete a sexual cycle is a common feature of hybrids (Hunter et al., 1996; Wolfe, 2015), including intra-species ones (Hou & Schacherer, 2016; Rogers et al.,

2018). Considering this, a hybridization scenario for the origin of *C. albicans* lineages could provide a plausible explanation for the origin of the inability of *C. albicans* to sporulate or undergo a standard sexual cycle. In this particular case, it could be hypothesized that the improper chromosome pairing after hybridization and consequent impossibility of completing meiosis contributed to the development of a parasexual cycle, an essential mechanism for *C. albicans* genomic plasticity. Alternatively, parasex might be a more ancient trait in the clade, predating the origin of the proposed hybridization. Supporting this is the finding that the closely related species *C. tropicalis* has also been shown to undergo a parasexual cycle under laboratory conditions (Seervai et al., 2013). If that is the case, hybridization between two lineages might have occurred through a sexual or parasexual cycle. Mating between different *Candida* species has been observed in the laboratory (Pujol et al., 2004), although it is unclear how widespread is this ability among *Candida*. Determining the exact mechanism of hybridization is beyond the scope of our study. However, based on our observations, we favor scenarios, such as standard sexual mating, in which two haploid cells fuse to form a diploid hybrid. Indeed, the finding that heterozygous regions are widespread across the genome and present in all chromosomes is at odds with expectations from parasexual crossing. In parasex, two diploid cells fuse to form an unstable tetraploid that quickly returns to a diploid state through concerted chromosomal loss. This process would rarely yield a chromosomal set composed of a copy from each parent for all the chromosomes, whereas this is what is expected by fusion of haploid cells. Although *C. albicans* belongs to a clade of diploid yeasts, many diploid yeast species form haploid cells to undergo mating. Furthermore, a viable mating-competent haploid state has been demonstrated for *C. albicans* (Hickman et al., 2013). We believe that further research is needed to clarify whether asexuality is a result or a facilitator of hybridization in this and other hybrid cases.

Our results raise once more the question of the importance of hybridization for the emergence of yeast pathogens (Mixão & Gabaldón, 2018) and pose the intriguing question of whether *C. albicans* ability to colonize and infect humans was an emerging phenotype enabled by this hybridization event. Indeed, hybridization is an important evolutionary mechanism that generates highly heterozygous genomes. This heterozygosity is often a source of

genomic plasticity that allows the emergence of new phenotypes or even adaptation to new niches. Many examples on different fungal species have showcased the relevance of hybridization for adaptation or diversification (Li et al., 2012; Smukowski Heil et al., 2017; Tusso et al., 2019; Zhang et al., 2020). From a clinical perspective, hybridization can be regarded as a source of new potentially pathogenic lineages (Mixão & Gabaldón, 2018). Many hybrids are becoming important agents of human infection, as it is the case of *Candida* or *Cryptococcus* species (Hagen et al., 2015; Mixão et al., 2019; Pryszcz et al., 2014, 2015; Samarasinghe et al., 2020; Schröder et al., 2016) with hybridization being also associated to a possible increase in virulence (Li et al., 2012). In a world where globalization and global warming are a reality promoting the expansion of certain species to locations where they have never been, the chances of new events of hybridization are possibly increasing. In this context, the impact of such events for public health should be regarded with some concern (King et al., 2015). In the particular case of the *Candida* clade, multiple hybridization events leading to the emergence of pathogenic lineages have been described (Mixão et al., 2019; Pryszcz et al., 2015; Schröder et al., 2016). This shows that this clade has some propensity to hybridize and contribute to the emergence of pathogens. The reason why this happens is difficult to be addressed. More studies should be performed to clarify this question and uncover the mechanisms of success of these pathogenic hybrid lineages.

8.5 Conclusion

This work assessed the origin of the high levels of heterozygosity in the important opportunistic pathogen *C. albicans*. We compared the genomic patterns of different strains and showed that *C. albicans* has a hybrid ancestor. This species is not the first hybrid lineage described in the *Candida* clade, but it is by far the most important one for the clinical setting. Why this clade has so many hybrid lineages is still unknown, but more studies should be performed trying to understand the particularities that make this clade so prone to hybridize. This finding raises once more the question about an apparent link between hybridization events and the emergence of pathogenic lineages. Therefore, future studies addressing the origins

of pathogenicity should consider the contribution of non-vertical evolution to this event.

8.6 Methods

NGS data selection

C. albicans paired-end reads used in this work are a subset of the data made publicly available under the BioProjects PRJNA432884 and PRJEB27862 (Bensasson et al., 2019; Ropars et al., 2018). Our sample was chosen based on different criteria. Specifically, we selected at least one strain from each SNP-based clade defined by Ropars et al., including *C. africana*, as this clade is highly homozygous (Ropars et al., 2018) and could correspond to a putative parental lineage. For clusters with more than one site of collection, one strain from each site was taken. In addition, the *C. albicans* type strain and two other environmental isolates were retrieved from (Bensasson et al., 2019). In the end, our sample consisted of a total of 61 *C. albicans* strains and eight *C. africana* (Additional file 2: Table S1).

In order to compare our results with other species, we retrieved raw reads of Illumina paired-end sequencing libraries from known hybrid and non-hybrid strains from diverse *Candida* species. As the representative of hybrid strains, we selected *C. orthopsilosis* MCO456 (BioProject PRJEB4430, SRA ERX295059); s425, s433, and s498 strains (BioProject PRJNA322245, SRA SRX1776098, SRX1776103, and SRX1776124); and *C. metapsilosis* CP367 (BioProject PRJEB1698, SRA ERX221928) (Pryszcz et al., 2014, 2015; Schröder et al., 2016). As the representative of non-hybrid strains, we selected *C. orthopsilosis* s428 (BioProject PRJNA322245, SRA SRX1776102), three *C. parapsilosis* strains (BioProjects PRJEB1685 and PRJNA326748, SRA ERX221039, ERX221044, and SRX1875155), and *C. tropicalis* ATCC200956 (BioProject PRJNA194439, SRA SRR868710) (Pryszcz et al., 2013; Schröder et al., 2016; Vincent et al., 2013).

Library preparation and genome sequencing

As we considered it important to compare *C. albicans* with the closely related species *C. dubliniensis* and *C. tropicalis*, we decided to sequence two strains from our lab collections, namely 60-13 (*C. dubliniensis*) and CSPO (*C. tropicalis*). Genomic DNA extraction was performed using the MasterPure Yeast DNA Purification Kit (Epicentre, USA) following the manufacturer's instructions. Briefly, cultures were grown in an orbital shaker overnight (200 rpm, 30°C) in 15 ml of YPD medium. Cells were harvested using 4.5 ml of each culture by centrifugation at maximum speed for 2 min, and then, they were lysed at 65°C for 15 min with 300 µl of yeast cell lysis solution (containing 1 µl of RNAse A). After being on ice for 5 min, 150 µl of MPC protein precipitation reagent was added into the samples, and they were centrifuged at 16,000*g* for 10 min to pellet the cellular debris. The supernatant was transferred to a new tube; DNA was precipitated using 100% cold ethanol and centrifuging the samples at 16,000*g*, 30 min, 4°C. The pellet was washed twice with 70% cold ethanol, and once the pellet was dried, the sample was resuspended in 100 µl of TE. All gDNA samples were cleaned to remove the remaining RNA using the Genomic DNA Clean & Concentrator kit (Epicentre) according to the manufacturer's instructions. Total DNA integrity and quantity of the samples were assessed by means of agarose gel, NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, USA), and Qubit dsDNA BR assay kit (Thermo Fisher Scientific).

Whole-genome sequencing was performed at the Genomics Unit from the Centre for Genomic Regulation (CRG) with a HiSeq2500 machine. Libraries were prepared using the NEBNext Ultra DNA Library Prep kit for Illumina (New England BioLabs, USA) according to the manufacturer's instructions. All reagents subsequently mentioned are from the NEBNext Ultra DNA Library Prep kit for Illumina if not specified otherwise. One microgram of gDNA was fragmented by ultrasonic acoustic energy in Covaris to a size of ~ 600 bp. After shearing, the ends of the DNA fragments were blunted with the End Prep Enzyme Mix, and then, NEBNext Adaptors for Illumina were ligated using the Blunt/TA Ligase Master Mix. The adaptor-ligated DNA was cleaned up using the MinElute PCR Purification kit (Qiagen, Germany), and a further size selection step was performed using an agarose gel. Size-selected

DNA was then purified using the QIAgen Gel Extraction Kit with MinElute columns (Qiagen), and library amplification was performed by PCR with the NEBNext Q5 Hot Start 2× PCR Master Mix and index primers (12–15 cycles). A further purification step was done using AMPure XP Beads (Agentcourt, USA). Final libraries were analyzed using Agilent DNA 1000 chip (Agilent) to estimate the quantity and check size distribution, and they were then quantified by qPCR using the KAPA Library Quantification Kit (KapaBiosystems, USA) prior to amplification with Illumina's cBot. Libraries were loaded and sequenced 2×125 bp on Illumina's HiSeq2500. Base calling was performed using the Illumina pipeline software. In multiplexed libraries, we used 6-bp indexes. De-convolution was performed using the CASAVA software (Illumina, USA).

Raw sequencing data analysis

Raw sequencing data was inspected with FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Paired-end reads were filtered for quality below 10 or 4 bp sliding-windows with average quality per base of 15 and for the presence of adapters with Trimmomatic v0.36 (Bolger et al., 2014). A minimum read size was set to 31 bp. To discard the possibility that higher quality thresholds would change our results, the trimming process was repeated for the five strains with lower coverage, using a minimum quality threshold of 28. After read mapping and variant calling (see the "Read mapping and variant calling" section), the main difference was the read depth of the called variants, which was lower in the stricter filter, decreasing the number of variants passing the filtration process (Additional file 12: Table S8). These results suggest that for this data, the quality threshold of 15 represents a good compromise between the read quality and the depth of coverage.

The K-mer Analysis Toolkit (KAT, (Mapleson et al., 2017)) was used to get the 27-mer frequency (default *k*-mer size) and GC content of each library. This program was also used to inspect the representation of each 27-mer in the respective reference genome. The genome assemblies used as reference were as follows: *C. albicans* assembly 22 haplotype A and B in separate (Muzzey et al.,

2013), *C. orthopsilosis* 90-125 ASM31587v1 (Riccombeni et al., 2012), *C. metapsilosis* chimeric genome assembly (Pryszcz et al., 2015), *C. parapsilosis* ASM18276v2 (Pryszcz et al., 2013), *C. tropicalis* ASM633v3 (Butler et al., 2009), and *C. dubliniensis* ASM2694v1 (Jackson et al., 2009).

Read mapping and variant calling

Read mapping of each sequencing library to the respective reference genome assembly was performed with BWA-MEM v0.7.15 (Li, 2013). It is important to note that for *C. albicans*, read mapping was performed in separate on haplotypes A and B. Picard integrated in GATK v4.0.2.1 (McKenna et al., 2010) was used to sort the resulting file by coordinate, as well as to mark duplicates, create the index file, and obtain the mapping statistics. The mapping results were visually inspected with IGV version 2.4.14 (Thorvaldsdóttir et al., 2013). Mapping coverage was determined with SAMtools v1.9 (Li et al., 2009).

SAMtools v1.9 (Li et al., 2009) and Picard integrated in GATK v4.0.2.1 (McKenna et al., 2010) were used to index the reference and create its dictionary, respectively, for posterior variant calling. GATK v4.0.2.1 (McKenna et al., 2010) was used to call variants with the tool HaplotypeCaller set with "--genotyping-mode DISCOVERY --standard-min-confidence-threshold-for-calling 30 -ploidy 2". The tool VariantFiltration of the same program was used to filter the vcf files with the following parameters: "-G-filter-name "heterozygous" -G-filter "isHet == 1" --filter-name "BadDepthofQualityFilter" -filter "DP <= 20 || QD < 2.0 || MQ < 40.0 || FS > 60.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0" --cluster-size 5 --cluster-window-size 20". In order to determine the number of SNPs/kb, a file containing only SNPs was generated with the SelectVariants tool. For this calculation, only positions in the reference with 20 or more reads were considered for the genome size, and these were determined with bedtools genomecov v2.25.0 (Quinlan & Hall, 2010).

Heterozygous and homozygous blocks definition

To determine for each highly heterozygous strain the presence of heterozygous and LOH blocks, we adapted the procedure validated by Pryszcz et al., (Pryszcz et al., 2015). Briefly, bedtools merge v2.25.0 (Quinlan & Hall, 2010) with a window of 100 bp was used to define heterozygous regions, and by opposite, LOH blocks would be all non-heterozygous regions in the genome. The minimum LOH and heterozygous block size was established at 100 bp. All regions that did not pass the requirements to be considered LOH or heterozygous blocks were classified as "undefined regions." No filter for coverage was applied due to the low coverage of some libraries. For hybrid strains, the current divergence between the parentals was calculated by dividing the number of heterozygous positions by the total size of heterozygous blocks.

We used a different method to define heterozygous blocks as compared to other studies because we consider that window-based approaches overestimate heterozygous block sizes, as window boundaries will rarely coincide with real heterozygous block boundaries. To confirm that this different approach (and not differences in variant calling) explains differences in levels of heterozygosity described in previous studies for the same strains, we repeated the analysis of SC5314 using the methodology described by Hirakawa et al., and Bensasson et al., (sliding-window approach), expecting to recover results similar to theirs (Bensasson et al., 2019; Hirakawa et al., 2015). As shown in Additional file 13: Table S9, applying Hirakawa et al.,'s method, we estimate ~80% heterozygosity for SC5314, which is similar to what they estimated (Hirakawa et al., 2015). In the case of Bensasson et al., although they do not indicate an estimation of heterozygosity, they report that the average of heterozygosity in windows with more than 0.1% SNPs is >0.4%, and we estimated it to be 0.5%, which is consistent (Bensasson et al., 2019). Furthermore, we could observe that depending on the window size, different estimations were made (Additional file 13: Table S9), with shorter windows estimating lower levels of heterozygosity, and apparently being more precise. This indicates that our results are not an artifact of variant calling.

Comparison of SNPs across different strains

If *C. albicans* is a hybrid, we would expect that the majority of heterozygous SNPs in heterozygous blocks would be shared by the different strains, as this would mean that they were present before they diverged. On the other hand, in LOH blocks, we would expect exactly the opposite, as the majority of heterozygous SNPs should correspond to new acquired mutations. To check this scenario, we compared the heterozygous and LOH blocks of four strains from different clades. Based on the phylogeny described by Ropars et al. (2018), we decided to consider three groups of strains, which worked as replicates of the analysis (Figure 8_2A). Each group was comprised of two strains from each side of the first node of divergence of *C. albicans* strains. To ensure that the results were not influenced by events of recombination between different clades, we only selected strains that did not present signs of admixture with other clades according to Ropars et al., (Ropars et al., 2018). The first group of strains comprised CEC4492 (clade 11) and SC5314 (clade 1) from one side, and CEC4497 (clade 4) and CEC5026 (clade 9) from the other side. The second one comprised CEC1426 (clade 11) and CEC3544 (clade 1) from one side, and CEC3716 (clade 4) and CEC3533 (clade 9) from the other side. Finally, the third group comprised CEC3601 (clade 11) and CEC3560 (clade 1) from one side, and CEC2021 (clade 2) and CEC3557 (clade 3) from the other side. For each group, we obtained the intersection of their LOH blocks and the intersection of their heterozygous blocks using bedtools intersect v2.25.0 (Quinlan & Hall, 2010). For each intersection, we inspected the heterozygous positions in each strain and compared them with the observed genotypes in the other three clades. For each position, we reconstructed the most parsimonious scenario by assessing all possible mutational paths and choosing the one with the lower number of inferred mutations. When this scenario pointed to a similar heterozygous genotype in the common ancestor of the four strains, this position was assigned as "ancestral" for that strain. In case it would point to a different genotype, this heterozygous position was assigned as "recent". When it was not possible to find a unique best scenario, the position was assigned as "ambiguous". We also performed this analysis considering only blocks with an intersection > 100 bp. Furthermore, another analysis of both LOH and heterozygous block intersections was performed separately for coding and non-coding regions. For that, bedtools

intersect v2.25.0 (Quinlan & Hall, 2010) was used to obtain for each intersection the blocks with at least one overlap with coding regions annotated for *C. albicans* assembly 22 and available at Candida Genome Database (Skrzypek et al., 2017). Detailed information on the size of the intersection and positions considered for analyses are detailed in Additional file 7: Table S5.

Phylogenetic analysis considering the two haplotypes

The phylogenetic analysis of *C. albicans* considering the two haplotypes was performed individually for each of the mentioned groups of strains. Thus, for each group, we selected the intersection of heterozygous blocks (comprising the two haplotypes), which were defined as described above. To make sure that in the final alignment haplotypes A and B blocks were correctly concatenated, only regions overlapping SC5314 heterozygous blocks were taken into consideration, because they are the only ones correctly phased in the genome assembly. Then, for each strain, the respective homozygous SNPs were substituted in the reference genome. To separate the two haplotypes, HapCUT2 (Edge et al., 2017) was used to phase the heterozygous variants. For each block, the most similar haplotype to SC5314 haplotype A was considered A, while the most distant one was considered B. Positions with INDELs in at least one of the strains were excluded. In the end, for each group, we had an alignment of the two haplotypes of each strain. A maximum likelihood tree representative of each alignment was obtained with RAxML v8.2.8 (Stamatakis, 2014), using the GTRCAT model and 1000 bootstraps. Midpoint rooting method was used to root the trees.

Comparison of SNPs between *C. albicans* and *C. africana* strains

To assess whether *C. africana* was one of *C. albicans* parental lineages, we compared the heterozygous positions of *C. albicans*, with the homozygous positions of *C. africana*. As *C. albicans* genome was phased based on SC5314 (Muzzey et al., 2013), proved in this work to be a hybrid strain, only heterozygous regions in this strain are expected to represent the two parental haplotypes in the reconstructed phased genome. This means we can only trust that

haplotype A of the reference genome corresponds always to the same parental in SC5314 heterozygous blocks. Therefore, for this analysis, we considered the intersection of these blocks with the heterozygous positions of each of the groups of *C. albicans* strains (check previous sections for details) and with the homozygous blocks defined for each *C. africana* strain. This analysis was performed independently for each group and each *C. africana* strain. Then, for each of these regions, we counted how many ancestral or recent positions identified in each of the four *C. albicans* strains of a given group (check previous sections for details) were shared (the same genotype was found in *C. albicans* and *C. africana*), or corresponded to a haplotype (haplotype A or B was present in both strains), or were undefined (none of the previous options was observed). Details on this analysis can be found in Additional file 11: Table S7.

8.7 Supplementary information

Supplementary information accompanies this paper at https://doi.org/10. 1186/s12915-020-00776-6.

8.8 Acknowledgements

The authors thank Dr. Emilia Gomez, for providing CSPO strain, and all Gabaldón's group, especially Susana Iraola, for laboratory work, and Marina Marcet-Houben, for the helpful discussions.

8.9 Authors' contributions

TG supervised the study. VM performed all the bioinformatics analysis. TG and VM wrote the manuscript. All authors read and approved the final manuscript.

8.10 Funding

This work was funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie

grant agreement no. H2020-MSCA-ITN-2014-642095. TG group also acknowledges support from the Spanish Ministry of Economy, Industry, and Competitiveness (MEIC) for the EMBL partnership, and grants "Centro de Excelencia Severo Ochoa" SEV-2012-0208, and BFU2015-67107 co-founded by European Regional Development Fund (ERDF); from the CERCA Program/Generalitat de Catalunya; from the Catalan Research Agency (AGAUR) SGR857; and grants from the European Union's Horizon 2020 research and innovation program under the grant agreements ERC-2016-724173, and MSCA- 747607. TG also receives support from an INB Grant (PT17/0009/0023 - ISCIII- SGEFI/ERDF).

8.11 Availability of data and materials

All data generated or analyzed during this study are included in this published article, its supplementary information files, and publicly available repositories. Specifically, the datasets generated during the current study are available in NCBI under the BioProject PRJNA555042. Data made publicly available by other projects are available in NCBI under the BioProject numbers PRJNA432884, PRJEB27862, PRJEB4430, PRJNA322245, PRJEB1698, PRJEB1685, PRJNA326748, PRJNA194439, PRJEA83665, PRJEA32889, PRJNA13675, and PRJEA34697. *C. albicans* assembly 22 is available in the Candida Genome Database.

9 Effect of drift, selection, and recombination on the evolution of genomes of hybrid yeast pathogens

Mixão, V., Ksiezopolska, E., Saus, E., Boekhout, T., Gacser, A., & Gabaldón, T. Effect of drift, selection, and recombination on the evolution of genomes of hybrid yeast pathogens. (In preparation)

9.1 Abstract

Hybrids are chimeric organisms with highly plastic, heterozygous genomes that may confer them with unique traits enabling them to adapt to new environments. However, most evolutionary theoretical frameworks predict that the high levels of heterozygosity present in hybrids from divergent parents are likely to result in numerous deleterious epistatic interactions. Under this scenario, selection is expected to favor recombination events resulting in loss of heterozygosity (LOH) affecting genes involved in such negative interactions. Nevertheless, it is so far unknown whether this phenomenon actually drives genomic evolution in natural populations of hybrids. To determine the balance between selection and drift in the evolution of LOH patterns in yeast hybrids, we analyzed genomic sequences from fifty-four hybrid strains of the pathogenic yeasts *Candida orthopsilosis* and *Candida metapsilosis*, which derived from at least five distinct natural hybridization events. Our results suggest that genetic drift is the prevailing force shaping LOH patterns in these hybrid genomes. Moreover, the observed LOH patterns suggest that these are likely not the result of continuous accumulation of sporadic events - as expected by mitotic repair of rare chromosomal breaks - but rather of sharp episodes involving many LOH events in a short period of time. We posit that escape from meiotic arrest provides a plausible explanation to such patterns.

9.2 Introduction

The cross between two diverged lineages originates a hybrid organism carrying a chimeric genome. As a result, hybrid genomes are highly heterozygous, with a genetic divergence between pairs of homeologous chromosomes initially equal to the divergence of the two parental species (Mixão & Gabaldón, 2018; Runemark et al., 2019). As previously suggested by Bateson, Dobzhansky and Muller (Bateson, 1909; Dobzhansky, 1934; Muller, 1942), this coexistence, not only of the two diverged genomes, but also of their respective transcripts and proteins, can originate incompatibilities that negatively impact cell function and compromise hybrid's fitness. Despite this fitness cost, if hybridization provides a key advantage in a given niche, hybrid lineages may survive (Gladieux et al., 2014). Under these circumstances natural selection is expected to favor loss of heterozygosity (LOH) in genomic regions comprising heterozygous genes involved in genomic incompatibilities (Mixão & Gabaldón, 2018; Runemark et al., 2019). Several processes can shape hybrid genomes, including the duplication or loss of chromosomes leading to chromosomal aneuploidies, gene loss, gene conversion, or whole genome duplication (Albalat & Cañestro, 2016; Marcet-Houben & Gabaldón, 2015; McGrath et al., 2014; Wertheimer et al., 2016; Wolfe, 2015). These mechanisms contribute to progressive LOH and promote genome stabilization by reducing the amount of heterozygosity and genomic incompatibilities.

Hybridization may occasionally lead to the emergence of new lineages or species, with some studies pointing to a role of hybridization in the emergence of new human pathogens (Mixão et al., 2019; Mixão & Gabaldón, 2020; Pryszcz et al., 2015). Indeed, recent genomic analyses have uncovered numerous hybrids among opportunistic fungal human pathogens (Hagen et al., 2015; Mixão et al., 2019; Mixao & Gabaldón, 2020; Pryszcz et al., 2014, 2015; Schröder et al., 2016). For instance, *Candida parapsilosis* species complex comprises three opportunistic pathogenic species, from which two, *Candida orthopsilosis* and *Candida metapsilosis*, have hybrid strains (Pryszcz et al., 2014, 2015; Schröder et al., 2016; Tavanti et al., 2005). All *C. metapsilosis* strains analyzed thus far are hybrids and all are inferred to descend from a single hybridization

event between two unknown lineages with 5% divergence at the nucleotide level (Pryszcz et al., 2015). In *C. orthopsilosis*, most analyzed strains are hybrids and the few known homozygous strains always correspond to the same parental lineage of these hybrids (Pryszcz et al., 2014; Schröder et al., 2016). In this case, at least four independent hybridization events have been inferred, all involving the same two lineages with 4.5% nucleotide divergence (Pryszcz et al., 2014; Schröder et al., 2016). This indicates a high propensity of this species complex to form hybrids that can successfully colonize and cause disease in humans. The fact that homozygous strains from the parental lineages are relatively less frequent or totally absent among clinical isolates prompted the hypothesis that hybridization events between non-pathogenic or lowly-pathogenic species led to the emergence of new hybrid lineages with increased virulence (Mixão & Gabaldón, 2018; Pryszcz et al., 2015).

Given the importance of hybrids not only for the clinics, but also for industry and biotechnology (Mixão & Gabaldón, 2018; Morales & Dujon, 2012), multiple studies have sought to understand the evolution of hybrid genomes, with a special focus on the possible sources of genomic incompatibilities and on the mechanisms of genome stabilization (Jhuang et al., 2017; Lancaster et al., 2019; Lee et al., 2008; Morard et al., 2020; Smukowski Heil et al., 2017). In this respect, suboptimal interactions between nuclear and mitochondrial components, or among ribosomal subunits when proteins are encoded by different parental genomes, have been proposed as important sources of genomic incompatibilities that compromise hybrids' fitness (Barreto & Burton, 2013; Jhuang et al., 2017; Lee et al., 2008). The availability of dozens of fully sequenced genomes of strains from independently formed hybrids in the *C. parapsilosis* clade provides a unique opportunity to study the genomic aftermath of hybridization and test some of the above-mentioned theoretical predictions. For instance, if the resolution of genomic incompatibilities is indeed a strong selective force, one would expect that LOH patterns among naturally occurring isolates are not randomly distributed along the genome, but rather enriched in regions comprising certain genes, such as those encoding ribosomal or mitochondrial proteins (Jhuang et al., 2017; Lee et al., 2008). To test this hypothesis and to gain insight into the evolutionary aftermath of hybridization at the genomic level, we

undertook a comparative analysis of 42 publicly available, and 17 newly sequenced strains from the *C. parapsilosis* clade.

9.3 Results

Genetic and geographical structure of hybrid clades, and the first environmental *C. metapsilosis* hybrid

To assess genome evolution following hybridization we analyzed publicly available genomic data for *C. orthopsilosis* and *C. metapsilosis* (Pryszcz et al., 2015; Schröder et al., 2016; Zhai et al., 2020), and sequenced additional strains from these species (see Material and Methods). We also produced an improved assembly for *C. metapsilosis* based on a hybrid approach combining long- and short-read sequencing technologies (see Material and Methods) and compared available reference genomes for both species to choose the best reference (File S1). In addition, we used as technical replicates re-sequencing data from the same strain obtained from two different laboratories to validate and improve our methodology to define LOH blocks (see Material and Methods, File S2).

The final dataset comprised 41 *C. orthopsilosis* and 18 *C. metapsilosis* strains, including the first sequenced *C. metapsilosis* strain isolated from an environmental source (Table S1, and Files S2, S3 and S4). Our results show that all *C. metapsilosis* and the majority of *C. orthopsilosis* isolates are hybrids, which is consistent with previous studies (Pryszcz et al., 2015; Schröder et al., 2016). The similarity in LOH patterns can reveal whether two strains likely result from the same hybridization event (Pryszcz et al., 2015). Overall, for the shared strains, pairwise comparisons of LOH patterns agreed with the *C. orthopsilosis* clades previously described (Schröder et al., 2016), except for clade 4 where the high level of heterozygosity and scarcity of LOH blocks challenges this type of comparisons (File S5). Based on this analysis, we assigned the eight newly sequenced *C. orthopsilosis* hybrid strains into existing clades (Table S1).

We next inferred a molecular phylogeny for *C. orthopsilosis* strains based on their polymorphisms (Figure 9_1a, see Material and

Methods). The phylogeny broadly supported previously defined clades, and the above suggested clade-adscription, and revealed that all new homozygous strains of *C. orthopsilosis* were close to s428, and therefore corresponded to the same, previously known, parental lineage (parental A). Notably, all clades have a broad geographical distribution, comprising isolation sites from at least two different continents (Figure 9_2). This broad geographical distribution is also apparent in the most heterozygous (and likely more recently formed) clade 4, which now comprises isolates from USA, Europe, and China.

Our sampling includes the first genome for an environmental isolate of *C. metapsilosis* (strain 11127). Our molecular phylogenetic analysis places this hybrid among clinical isolates. This finding would support the previously proposed scenario of an environmental origin of *C. metapsilosis* clade in which hybrids, contrarily to their parental species, have the ability to colonize and infect humans (Pryszcz et al., 2015). Other *C. metapsilosis* strains have been reported from the environment, including samples from rice plants and, consistent with the source of our isolate, marine locations (Kaewkrajay et al., 2020; Khunnamwong et al., 2018; Xu et al., 2014). However, our analysis of deposited ITS data from these other environmental samples revealed that they likely represent different, unidentified species from the *Lodderomyces* clade (see Figure S1).

Initial analysis of *C. metapsilosis* strains suggested a single hybrid origin for all of them (Pryszcz et al., 2015). However, a recent study considering a new strain of *C. metapsilosis* (MSK446, included in our dataset) showed that it is phylogenetically distant from the previously described clade (Zhai et al., 2020). Our analysis of LOH blocks overlap is inconclusive, not allowing the confirmation of the number of hybridization events for this species (File S5). Nevertheless, the molecular phylogeny of *C. metapsilosis* strains (Figure 9_1b) confirmed that all of them, except CBS2916 and MSK446, form a tight clade (clade 1). CBS 2916 and MSK446 form a distantly related clade 2. The divergence between these two clades is comparable to the one observed for the different hybridization events in *C. orthopsilosis*. Previous analyses had reported a LOH event resulting in a partially overwritten *MTLa* idiomorph with *MTL*α (Pryszcz et al., 2015). We found that this event was also

present in strains of clade 2, albeit with slightly different boundaries (Figure S2).

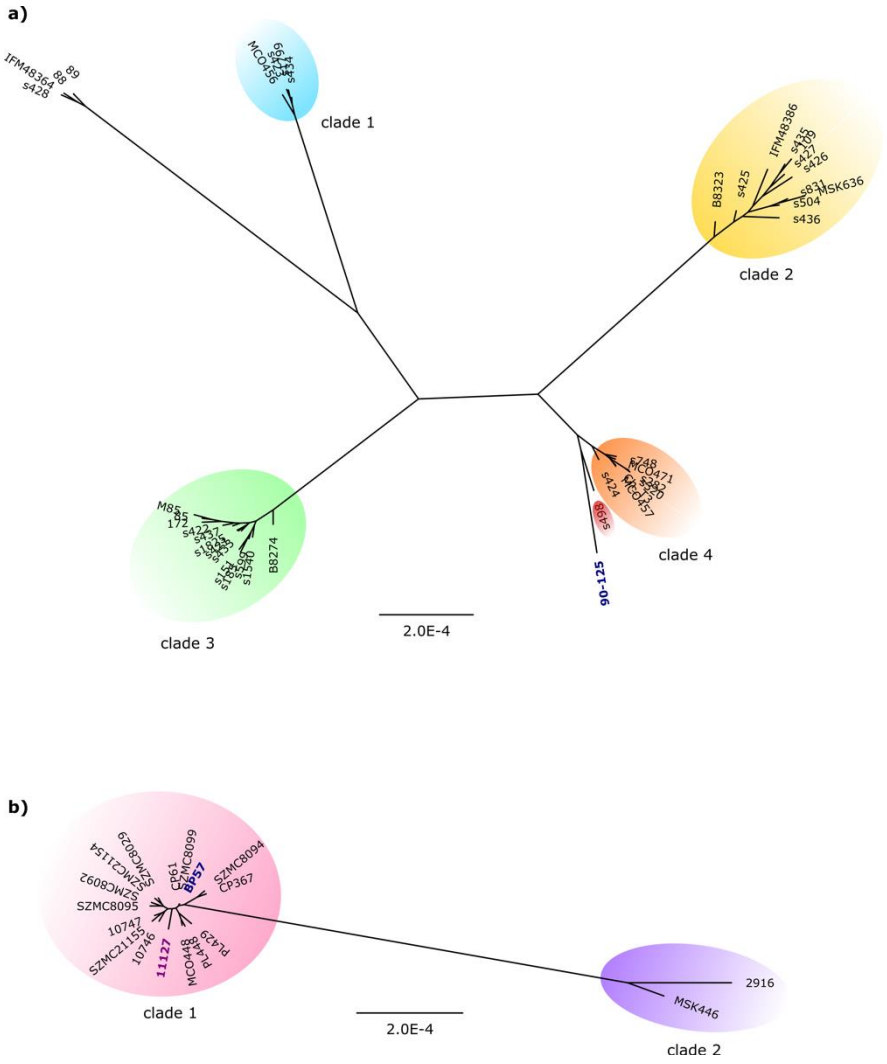

Figure 9_1. Maximum Likelihood phylogenetic trees. **a)** *C. orthopsilosis* tree, with clade 1 marked in blue, clade 2 in yellow, clade 3 in green, and clade 4 in orange. Due to its dubious clade adscription, s498 is marked in red. **b)** *C. metapsilosis* tree, with clade 1 marked in pink and clade 2 in purple. The environmental isolate is written with purple color. Both reference genomes are in blue.

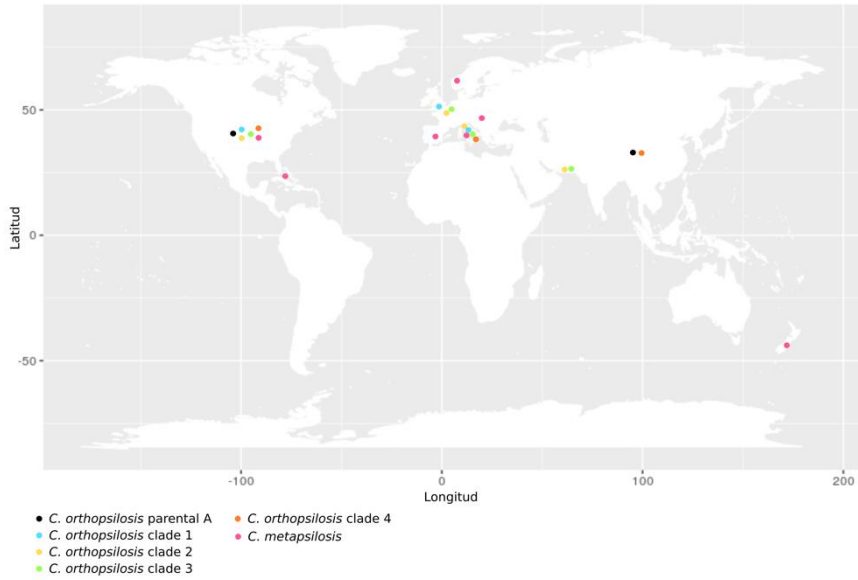

Figure 9_2. Geographical distribution of the strains analysed in this project. Circles, with colors representing different clades, are placed on the country where at least one sample of a given clade was isolated.

To investigate whether *C. metapsilosis* clades 1 and 2 correspond to different hybridization events, we assessed the number of shared LOH blocks, with the exact same boundaries, in all *C. metapsilosis* strains. We found two LOH blocks with an approximate length of 1 kb, which are shared by all *C. metapsilosis* strains, including CBS 2916 and MSK446. For comparison, distinct *C. orthopsilosis* clades do not share blocks with more than 1 kb with the exact same boundaries (File S6). The shared LOH blocks partially overlap *MTC5* and *BPH1* genes. These genes are highly heterozygous in some *C. orthopsilosis* strains, indicating that their sequences are not particularly conserved in the clade. Thus, a scenario involving two hybridization events would imply either the occurrence of two convergent events (plus a very similar one in the *MAT* locus) or the exchange of these genetic regions. We consider that with the current data at hand, a shared hybridization event followed by early separation of the two clades is the most parsimonious scenario, as it does not imply any convergent LOH event. This scenario would

involve a very early LOH event in the *MAT* locus followed by the separation of the two clades, with subsequent accumulation of independent LOH events, including the extension of LOH in the *MAT* locus of clade 2. Nevertheless, more studies are necessary to clarify this scenario.

Lack of strong functional trends in LOH patterns

To assess the strength of selection in shaping LOH patterns, we used the inferred LOH blocks to define the level of homozygosity per gene and strain, as the fraction of that gene covered by LOH blocks (see Material and Methods). For all clades, except *C. orthopsilosis* clade 1, there was clearly a fraction of genes with higher levels of LOH than others (File S7). We tested whether the fraction of shared homozygous (100% LOH) or heterozygous (0% LOH) genes between pairs of strains from different clades was different from random expectation and found that they are more congruent than randomly expected ($p < 0.01$, see File S8). This indicates that patterns of LOH in independent clades share some level of convergence that would not be expected from random occurrence, and thus provide evidence for the existence of some constraints. However, considering a minimum LOH block size of 100 bp, there was not a single gene completely homozygous or heterozygous in all the studied strains. When increasing this threshold to 1 kb, five genes were in heterozygous regions in all the studied strains, namely *TFA2* (required for promoter clearance by RNA polymerase), *NTF2* (essential in nucleocytoplasmic transport), *AIM21* (involved in mitochondrial migration along actin filaments), *QCR7* (component of the ubiquinol-cytochrome c reductase complex), and *DIC1* (mitochondrial carrier family). The first three and the last two genes form, respectively, two blocks of contiguous genes. We evaluated the behavior of these genes in hybrid strains of the distantly related *C. inconspicua* (Mixão et al., 2019) and found that *NTF2* and *QCR7* were in heterozygous regions in all hybrid isolates from this species. Although our results are inconclusive, this striking coincidence suggests that the heterozygosity of these genes is constrained in hybrids from *Candida* spp..

We next tested whether LOH blocks were preferentially covering genes performing certain functions. To do so, we functionally

annotated the genes and performed enrichment analyses for each independent hybridization event (see Material and Methods). To discard the possibility that highly conserved genes could be influencing our results, we limited our analysis to genes with at least one non-synonymous variant in at least one strain. While for *C. orthopsilosis* we did not find any particular enrichment, for *C. metapsilosis* we found that genes with more than 50% overlap with LOH blocks were enriched in cell-wall and cell surface-related functions, but only using a LOH block threshold of 100bp. Overall, these results show that, although the LOH patterns are not random, there is a lack of strong functional selection, suggesting that drift might be the prevailing force in shaping heterozygosity patterns.

Mitochondrial inheritance and LOH in mtDNA-interacting Pentatricopeptide repeat proteins (PPR)

The fact that different hybrid clades in *C. orthopsilosis* have retained mitochondria from the different parental lineages provided us with a unique opportunity to assess whether mito-nuclear incompatibilities are driving LOH patterns in yeast hybrids. Indeed, an earlier work determined that *C. orthopsilosis* clades 1 and 3 inherited the mitochondrial genome of parental A (corresponding to the reference genome), while clade 2 presented mitochondria from parental B, and clade 4 had a recombinant mitochondrion (Schröder et al., 2016). We here confirmed these results and determined the origin of the mitochondria of *C. metapsilosis* clade 2 as the same as clade 1.

In a scenario of strong mito-nuclear incompatibility, we would expect the direction of LOH (i.e. which sub-genome is retained) in mtDNA interacting proteins to follow the same direction of mitochondrial inheritance (i.e. same parental genome would be retained to minimize negative epistatic interactions). Earlier works have linked nuclear-encoded mitochondrial proteins (e.g. PPR) to incompatibilities in hybrid yeast species (Jhuang et al., 2017). We analyzed in detail thirteen genes predicted to encode PPR proteins in *C. metapsilosis* and *C. orthopsilosis* (File S9). Our results showed that, except for *C. metapsilosis* and *C. orthopsilosis* clade 4, most PPR proteins underwent full LOH in the majority of the strains. However, in the same *C. orthopsilosis* clade different PPRs lost their heterozygosity towards different parentals, and even the same PPR

underwent LOH towards different parentals in clades 1 and 3 (which retained the same mitochondrial background). These results indicate that the direction of LOH in PPR proteins does not follow the mitochondrial inheritance in *C. orthopsilosis* hybrids (File S9). Considering that all these strains are well succeeded and some of them have been kept in the laboratory for years, these results can possibly indicate that homozygosity in these genes is important, but that the specific retained allele is not so relevant for the hybrid survival. This is in agreement with what was previously suggested for the nuclear genome of these species (Pryszcz et al., 2014, 2015; Schröder et al., 2016), and with a recent analysis of *Saccharomyces* hybrids that revealed that the overall direction of LOH is not correlated with direction of mitochondrial inheritance (Langdon et al., 2019).

We analyzed in more detail two PPR genes for which experiments in other yeasts have proven a role in hybrid incompatibilities. For instance, *Aep3* was previously shown to be responsible for incompatibilities in *Saccharomyces cerevisiae* x *Saccharomyces pastorianus* hybrids (Jhuang et al., 2017). We found that this gene is 100% homozygous in all *C. orthopsilosis* strains, except s424 and MCO471, and the same pattern was observed in all *C. metapsilosis* strains. This suggests a selection for LOH in this gene, but also its non-essentiality. Furthermore, *Ccm1* which was previously suggested to be a source of mito-nuclear incompatibilities in *S. cerevisiae* x *Saccharomyces bayanus* hybrids (Jhuang et al., 2017) was found to be always within LOH blocks in all *C. metapsilosis* clade 1, and all *C. orthopsilosis* clade 2. As all clades of *C. orthopsilosis* were originated from the cross of the same two lineages, thus presenting at the time of hybridization similar nuclear content, only differing in the mitochondrial genome, we hypothesize that, contrarily to what was observed for *S. cerevisiae* x *S. bayanus* hybrids (Jhuang et al., 2017), in *C. orthopsilosis* hybrids, the incompatibility related to *Ccm1* is asymmetric, i.e. only occurs if parental B mitochondria is retained. The same authors showed that the strength of the incompatibility was dependent on carbon source. Therefore, an alternative scenario is that *C. metapsilosis* clade 1 and *C. orthopsilosis* clade 2 were under similar stresses during part of their evolution, selecting LOH in similar genes, like *Ccm1*. This last scenario would justify why despite having the same mitochondrial mitotype, this gene is not homozygous in *C. metapsilosis* clade 2.

LOH does not linearly accumulate with time

A key aspect to understand the evolution of LOH patterns is to discern the molecular mechanisms by which they are originated. LOH are thought to result from mitotic repair of chromosomal breaks and thus the relative level of LOH in hybrid strains is often taken as a proxy for the relative age of different hybridization events (Schröder et al., 2016). However, it is so far unknown what is the rate at which LOH blocks accumulate, and whether they accumulate linearly with time. We tested this idea by comparing the level of accumulated LOH with the amount of heterozygous mutations within shared LOH blocks in a given clade, as both mutational events must have postdated the divergence of the clade, and they should have occurred in the same window of time for each clade (see Material and Methods). *C. orthopsilosis* strain s498 was excluded from the analysis due to its dubious clade-adscription (see above, Table S1).

In this analysis, we estimated that each *C. metapsilosis* strain of clade 1 accumulated on average 0.01 heterozygous SNPs per kilo-base, and *C. orthopsilosis* clades 1, 2, 3 and 4 accumulated 0.07, 0.04, 0.004 and 0.07 heterozygous SNPs per kilo-base, respectively (File S10). Considering that *C. metapsilosis* strains of clade 2 were isolated in nature with 60 years of difference (CBS 2916 in 1954 and MSK446 in 2014), we decided to use their average accumulated heterozygous SNPs per kb as a rate to estimate the divergence time of the other clades. Our estimations point to approximately 28, 319, 178, 20, and 361 years of divergence for *C. metapsilosis* and *C. orthopsilosis* clades 1 to 4, respectively. These estimates are a rough proxy for the divergence time of these clades, and depend not only on the dataset, but above all on values calculated based on two strains which have been evolving in different environments (CBS 2916 in collection for 60 years and MSK446 recently isolated from human body). However, we consider that they are an appropriate proxy for comparing the relative age of the different clades, assuming they have similar mutation rates and generation times. Importantly, we did not observe any significant correlation between the estimated divergence times (or the mutational load) and the amount of LOH acquired after their divergence (Spearman rho = 0.22492, p = 0.53212), even when removing *C. metapsilosis* from the analysis (Spearman rho = 0.2381, p = 0.57016). We consider that these results

indicate that LOH does not accumulate with time in a linear way. Hence, we propose that other factors, besides time, are influencing the occurrence of LOH in hybrids of *C. parapsilosis* species complex.

C. orthopsilosis LOH patterns cannot be exclusively explained by mitotic recombination

Initial comparative genomics analyses of the first reported hybrid strain in *C. orthopsilosis* suggested that meiotic, in addition to mitotic, recombination could be in part responsible for the observed LOH patterns (Pryszcz et al., 2014). At that time, only two hybrid strains (MCO456 and AY2) were available. However, although the presence of alternative mating types in these hybrids suggests they originate through sexual crosses, meiosis has so far not been observed in the *C. parapsilosis* species complex (Butler et al., 2009; Pryszcz et al., 2014). Here, taking advantage of the larger number of hybrid strains comprising our dataset and the improved LOH definition procedure, we decided to revisit this hypothesis, as meiotic recombination could be a plausible explanation for the observed patterns of LOH.

Although inter-hom(e)ologous chromosomes recombination can occur during mitosis (Symington et al., 2014), most part of the chromosomal breaks are solved with sister chromatids (Bzymek et al., 2010; Kadyk & Hartwell, 1992). In contrast, during meiosis, inter-hom(e)ologous chromosomes recombination is an essential event, and therefore it is expected to occur at a higher rate (Dayani et al., 2011; Petronczki et al., 2003). In heterozygous genomes inter-hom(e)ologous chromosomes recombination may result in LOH. In our analysis we detected at least 262 LOH blocks longer than 1 kb in each *C. orthopsilosis* hybrid strain, a number that increased to 4,124 blocks when considering blocks longer than 100 bp. Furthermore, taking advantage that *C. orthopsilosis* reference genome corresponded to one of its hybrid parental lineages (Schröder et al., 2016), we screened the different strains for the presence of recombination events in heterozygous regions as well (see Material and Methods). We found evidence for recombination in all strains of clades 1 and 2 (except s427, s436 and s504), suggesting that the number of recombination events is even higher

than what we can detect by analyzing LOH. Therefore, the high number of recombination events is hardly explained by supposedly low mitotic recombination rates. A low rate of mitotic events leading to new LOH blocks is supported by the absence of differences in LOH blocks between the two isolates of strain s424 sequenced in this study, and years ago by another lab (see Material and Methods and File S11). Therefore, we reconsidered the possibility that hybrids in the *C. parapsilosis* clade are able to enter meiosis.

Previous studies have reported a positive correlation between the meiotic recombination rate and the chromosome size in *S. cerevisiae* and *Lachancea kluyveri* (Brion et al., 2017; Yin & Petes, 2013). Therefore, if *C. orthopsilosis* hybrids were able to undergo meiosis, one would expect a negative correlation between LOH and chromosome sizes, and correspondingly a positive correlation between the number of LOH events and chromosome size. Indeed, similarly to what was previously observed for MCO456 (Pryszcz et al., 2014), our results revealed that such correlations were statistically significant for all strains of clade 1, except s1799, which presented similar trends, but was not statistically significant (Table 9_1 and File S12). Besides clade 1, the strain s498 presented a significant negative correlation between LOH block size and chromosome size (Spearman rho = -0.7857, p = 0.02793) and positive correlation between the number of LOH events and chromosome size (Spearman rho = 0.7619, p = 0.03676). Therefore, some *C. orthopsilosis* hybrid strains present patterns compatible with footprints of meiotic recombination.

We set out to test this hypothesis experimentally following the same approach as Guitard et al. (2015), who reported sporulation in a *C. inconspicua* strain which was later proven to be a hybrid (Mixão et al., 2019) (see Material and Methods). Our attempts were, however, unsuccessful indicating either inability to undergo meiosis or failure to trigger that process in our experimental conditions.

Table 9_1. Spearman correlation test between the average number/size of LOH and chromosome size for the strains of *C. orthopsilosis* clade 1 and the strain s498. Spearman rho is indicated for each comparison, and the respective p-value is between parentheses.

Strain	LOH size vs. chromosome size	LOH number vs. chromosome size
MCO456	-0.7619048 (0.03676)	0.7619048 (0.03676)
s1799	-0.6428571 (0.09618)	0.7142857 (0.05759)
s423	-0.7619048 (0.03676)	0.7857143 (0.02793)
s434	-0.7619048 (0.03676)	0.7619048 (0.03676)
s498	-0.7857143 (0.02793)	0.7619048 (0.03676)

9.4 Discussion

Here, we analyzed the genomic aftermath of hybridization in the *C. parapsilosis* species complex, using improved assemblies and algorithms, and a comprehensive collection of 59 sequenced strains from at least 5 different hybridization events. This set includes 18 newly sequenced strains collected in an effort to find new hybridization events and the potential parental lineages of these hybrids. Three new strains of *C. orthopsilosis* were found to be homozygous and correspond to the already known parental A of these hybrids (Schröder et al., 2016). The absence of the unknown *C. orthopsilosis* parental B or any of the two unknown *C. metapsilosis* parents in the extended dataset reinforces the previously proposed idea (Mixão & Gabaldón, 2018) that the hybrids strains are over-represented across clinical isolates, which in turn suggests a higher capacity to infect humans as compared to their homozygous parentals.

Our new sampling comprises the first sequenced environmental isolate of *C. metapsilosis*, which we prove to be a hybrid. Analysis of deposited ITS sequences from other environmental samples did not allow us to discard the possibility of contamination but suggests that members of *C. parapsilosis* species complex can inhabit marine environments. This supports the idea that the parental lineages are possibly non-pathogenic, and that a (likely environmental)

hybridization event led to the emergence of a lineage able to more efficiently colonize (and infect) humans (Mixão & Gabaldón, 2018; Pryszcz et al., 2015). Considering the increasing number of hybrid pathogens being described (Mixão & Gabaldón, 2018), this process should be regarded with particular concern in a context of climate change and globalization promoting the contact between divergent lineages (King et al., 2015).

Hybrid genomes evolve through LOH which results in distinct patterns that differentiate strains and hybrid clades. We found that the number of overlap of homozygous and heterozygous regions among these hybrids is larger than what is expected to occur by chance and we found an extreme example of two genes that were always heterozygous not only in all the hybrid strains of the *C. parapsilosis* clade, but also in hybrids of the distantly related *C. inconspicua*. Moreover, similarly to what was previously described for other *Saccharomyces* hybrids (Jhuang et al., 2017), *Aep3* and *Ccm1* are also a likely source of genomic incompatibilities in hybrids of the *C. parapsilosis* clade. However, strains with similar nuclear and mitochondrial backgrounds at their origin (clades 1 and 3 of *C. orthopsilosis*) followed a different LOH direction in these genes, suggesting that the absence of high heterozygosity, but not the retention of a specific haplotype, is the target of selection. However, we found very few instances of functional enrichment in genes preferentially covered by LOH regions. In the case of *C. metapsilosis*, considering very relaxed thresholds, we found that cell-wall composition might have been under selection for reducing heterozygosity. Although this result disappeared when considering more strict thresholds, we speculate that perhaps the adaptation to a new environment (like the human body) could select for homozygosity in cell-wall composition genes. Altogether, these results indicate that, although we found some evidence for Bateson-Dobzhansky-Muller incompatibilities, they are apparently not widespread or too weak to drive genome-wide convergent functional patterns in independently formed hybrids, and thus genetic drift might be the prevailing force shaping LOH patterns. In addition, it is important to remind that, in the presence of non-selective constraints (such as regions prone to recombination) non-random distributions of LOH across genes might also emerge. Alternatively, functional enrichment tests are not sufficiently fine grained to detect the target of selection, or characteristics other than function, of the

genes might be the target of selection. In this respect, a recent study found that genes with allele specific expression were less likely to undergo LOH in *C. orthopsilosis* strains (Hovhannisyan et al., 2020), which can suggest that selection at the transcriptional level might be playing a role in retaining heterozygosity. Further transcriptomic analysis using a larger sample size might help to further clarify hybrid genome evolution.

The balance between selection and drift is influenced by different factors, including the size of the population, and the mechanisms by which genetic variability appears. Small population sizes and the existence of strong population bottlenecks limit the power of selection and favor genetic drift. We consider that this might have been an important factor contributing to the apparent weak signal for selection found in LOH patterns across *C. parapsilosis* hybrids. In addition, selection acts on entire genotypes and we consider that the nature of how LOH patterns are generated, in contrast to point mutations, also limits the power of selection to favor homozygosity at particular loci. In fact, a single LOH event can affect several genes and dozens of heterozygous loci at the same time, and the advantage of homozygosity at certain loci might be compensated by deleterious effects of losing some alleles at other loci. Even if selection does efficiently select particular homozygous loci, our ability to detect this selection will be hampered by the fact that this will be accompanied with many "passenger" genes that concomitantly became homozygous. These effects would be exacerbated if LOH occurs in an episodic manner, affecting multiple, scattered genomic areas in a short period of time. Our results indeed support an episodic nature of widespread LOH accumulation. Firstly, the finding of no single LOH difference in two independent isolates from the same strain maintained at different labs suggests that LOH patterns are stable, at least over hundreds of mitotic divisions. Secondly, contrary to the expectation for a constant rate of accumulation of LOH through mitotic divisions, we found no correlation between accumulation of LOH and accumulation of point mutations. These findings are compatible with an episodic nature of LOH accumulation, with periods of widespread LOH followed by long periods of stasis. A previous study showed that the occurrence of LOH is often related to stress response (Rosenberg, 2011), so we hypothesize that perhaps stress is one of the triggers of LOH in these hybrids and a major factor influencing their genomic features. This

would explain the absence of new LOH events in strains kept for years in the lab.

Additional observations led us to seriously consider the existence of meiotic recombination, at least in some *C. orthopsilosis* strains, as previously proposed (Pryszcz et al., 2014). First of all, the high amount of LOH blocks detected in the hybrids, together with the above-mentioned apparent non-linearity of their accumulation, are highly suggestive that recombination between inter-homeologous chromosomes might be induced by meiosis. As meiosis is generally induced by stress, this idea is compatible with the stress induced LOH hypothesis mentioned above. In support of meiotic breaks having a role in the formation of LOH, we found, in some strains, correlations between length of the chromosomes and size of LOH blocks, which would be unexpected from sporadic mitotic events. However, similarly to previous experimental attempts (Butler et al., 2009; Pryszcz et al., 2014), we failed to induce sporulation in the lab.

Recently, alternative mechanisms that promote inter-hom(e)ologous chromosomes recombination at a rate that is higher than what is expected for mitosis have been described. One of such mechanisms is parameiosis (Anderson et al., 2019), which was recently described in the opportunistic pathogen *C. albicans* (shown to be an evolved hybrid (Mixao & Gabaldón, 2020)). After parasex (mating of two *C. albicans* diploid cells) the tetraploid progeny tries to restore the diploid state through concerted chromosome loss (Berman & Hadany, 2012; Forche et al., 2008). This last process is what represents parameiosis, because it involves genes that are similar to those involved in meiotic recombination, and recombination occurs at very high rates (Anderson et al., 2019). Furthermore, this process seems to generate patterns of recombination compatible with the positive correlation between the number of recombination events and chromosome size that we observed for *C. orthopsilosis* hybrids (Anderson et al., 2019). Considering the failure to induce sporulation in *C. orthopsilosis*, we first hypothesized that parameiosis could be the underlying process of the LOH patterns we observed. However, such a mechanism would imply multiple rounds of polyploidization followed by chromosome loss. In such a scenario, we would not expect hybrids from the same clade to have so similar patterns, so we consider that it is improbable that parameiosis occured in *C. orthopsilosis* hybrids.

Another possible mechanism is the return to growth (RTG) after incomplete meiosis. This mechanism was shown for *S. cerevisiae* hybrids, which are able to enter meiosis, experience genomic recombination and return to growth without progressing to sporulation (Laureau et al., 2016; Simchen, 2009). RTG and meiosis involve inter-homeologous chromosomes recombination with a similar rate but differ in the way molecule junctions are solved (Laureau et al., 2016; Dayani et al., 2011). This mechanism could explain the high amount of recombination events detected in *C. orthopsilosis* hybrids without the need of sporulation and would reconcile our genomic and experimental findings. Thus, we hypothesize that this or a similar process is responsible, at least in part, for the observed genomic patterns in *C. orthopsilosis*, and possibly in *C. metapsilosis* as well. However, additional studies should be carried out to confirm or refute this hypothesis, and further clarify the mechanisms underlying LOH in *Candida* hybrids.

9.5 Material and Methods

Genomic DNA sequencing

A modified protocol from the MasterPure™ Yeast DNA Purification Kit was used to extract the DNA. In brief, samples were grown overnight in liquid YPD at 30˚C. Cells were pelleted and lysed with RNAse treatment at 65˚C for 15 min. After 5 min of cooling down on ice, samples were purified by the kit reagent by mixing, centrifugation and removal of the debris as described in the kit protocol. Further, samples were left at -20˚C with absolute ethanol for at least 2 h after which the DNA was precipitated for 30 min at 4˚C. The pellet was washed in 70% ethanol and left to dry. TE buffer was used to resuspend the DNA. Genomic DNA Clean & Concentrator kit was used for the final purification.

Whole-genome sequencing was performed at the Genomics Unit from CRG. As the samples were not sequenced all at the same time, but rather in two groups at different timepoints (group 1: 10746, 10747, 11127, 2916, SZMC21155, SZMC8099 and Ch_T3; group 2: 109, 172, 85, 88, 89, IFM48364, IFM48386, MCO457, MCO471),

the protocol was not always the same. Differences in protocol are mentioned across the methods description.

Libraries were prepared using the NEBNext® DNA Library Prep Reagent Set for Illumina® kit (New England BioLabs) according to the manufacturer's protocol. Briefly, 1 µg of gDNA was fragmented by nebulization in Covaris to approximately 600 bp and subjected to end repair, addition of "A" bases to 3′ ends and ligation of Truseq adapters. All purification steps were performed using Qiagen PCR purification columns (Qiagen) for group 1 and AMPure XP beads (Beckman Coulter) for group 2. Library size selection was done with 2% low-range agarose gels. Fragments with average insert size of 700 bp (for the group 1) and 665 bp (for the group 2) were cut from the gel, and DNA was extracted using QIAquick Gel extraction kit (Qiagen) and eluted in 30 µl EB. 10 µl of adapter-ligated size-selected DNA were used for library amplification by PCR using the Truseq Illumina primers. Final libraries were analyzed using Agilent DNA 1000 chip to estimate the quantity and check size distribution and were then quantified by qPCR using the KAPA Library Quantification Kit (KapaBiosystems, ref. KK4835) prior to amplification with Illumina's cBot. Libraries were loaded at a concentration of 2 pM onto the flow cell and were sequenced 2 x 125 bp on Illumina's HiSeq 2500.

Raw sequencing reads generated for this study can be found in SRA under the BioProject number PRJNA520893. The remaining genome sequencing libraries used for this work were retrieved from the BioProjects PRJEB4430, PRJEB1698, PRJNA322245 and PRJNA579121 (Zhai et al., 2020; Schröder et al., 2016; Pryszcz et al., 2014, 2015).

Genome assembly of *C. metapsilosis*

To obtain a better genome assembly for *C. metapsilosis*, we used a pipeline that combines both short and long-read assemblers to assemble sequencing data of the strain BP57 (Bastos et al., 2020). Briefly, BP57 illumina reads were filtered and trimmed with Trimmomatic v0.36 (Bolger et al., 2014) and assembled with Platanus v1.2.4 (Kajitani et al., 2014). Nanopore reads were corrected with Canu (Koren et al., 2017) and assembled with

DBG2OLC (v20180222) (Ye et al., 2016) using Platanus assembly, MaSurCA v3.3.0 (Zimin et al., 2013), and WTDBG2 v2.1 (Ruan & Li, 2020). Ragout v2.2 (Kolmogorov et al., 2014) was used for scaffolding using DBG2OLC, WTDBG2 and MaSurCA assemblies. Assembly correction was performed with Pilon v1.22 (Walker et al., 2014). The quality of each of the assemblies was assessed with Quast v4.5 (Gurevich et al., 2013) and K-mer Analysis Toolkit v2.4.1 (KAT, (Mapleson et al., 2017)). In the end, the assembly generated by Ragout was the one with the best results and used for downstream analysis (check the "Results" section and File S1). Augustus v3.5 (Stanke & Morgenstern, 2005) was used for genome annotation, using *Candida albicans* as model species. Assembly completeness was assessed with BUSCO v3 (Waterhouse et al., 2018).

Read mapping and variant calling

All paired-end Illumina libraries of the 18 *C. metapsilosis* strains and 41 *C. orthopsilosis* strains (Table S1) were inspected with FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed and filtered with Trimmomatic v0.36 (Bolger et al., 2014). Filtered reads were mapped on the respective genome references (*C. metapsilosis* BP57 assembly and *C. orthopsilosis* 90-125 assembly (Riccombeni et al., 2012)) with BWA-MEM v0.7.15 (Li, 2013). Picard integrated in GATK v4.0.2.1 (McKenna et al., 2010) was used to sort the resultant file by coordinate, as well as, to mark duplicates, create the index file, and obtain the mapping statistics. The mapping was inspected with IGV version 2.0.30 (Thorvaldsdottir et al., 2013). Mapping coverage was determined with SAMtools v1.9 (Li et al., 2009). Similar approach was used to align the genomic reads on the respective reference mitochondrial genome (accession numbers: JQ062879.1 for *C. metapsilosis* and NC_006972.1 for *C. orthopsilosis*). It is important to mention that in the case of the recently sequenced isolates described by (Zhai et al., 2020), multiple sequencing data was available for *C. metapsilosis* and *C. orthopsilosis*. However, as confirmed by the phylogenies provided by the same authors, all of them corresponded to the same strain, and therefore we have selected only one of each to include in our dataset (MSK446 in the case of *C. metapsilosis* and MSK636 in the case of *C. orthopsilosis*).

Samtools v1.9 (Li et al., 2009) and Picard integrated in GATK v4.0.2.1 (McKenna et al., 2010) were used to index the reference and create its dictionary, respectively, for posterior variant calling. GATK v4.0.2.1 (McKenna et al., 2010) was used to call variants with the tool HaplotypeCaller set with --genotyping_mode DISCOVERY -stand_emit_conf 10 -stand_call_conf 30 -ploidy 2 -nct 8. The tool VariantFiltration of the same program was used to filter the vcf files with the following parameters: --clusterSize 5 --clusterWindowSize 20 --genotypeFilterName "heterozygous" --genotypeFilterExpression "isHet == 1" --filterName "bad_quality" -filter "QD < 2.0 || MQ < 40 || FS > 60.0 || HaplotypeScore > 13.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0" --filterExpression "DP <= 30" --filterName "DepthofQuality". In order to determine the number of SNPs/kb, a file containing only SNPs was generated with the SelectVariants tool. Moreover, for this calculation only positions in the reference with more than thirty reads were considered for the genome size, and these were determined with bedtools genomecov v2.25.0 (Quinlan & Hall, 2010).

LOH blocks and gene homozygosity definition

Taking advantage of the presence of more than one library for the same strain in our dataset (s424), the reproducibility of the procedure for LOH blocks definition applied and validated by Pryszcz et al. (2015) was tested. Briefly, we determined LOH blocks for the different libraries of the same strain by filtering out regions with < 0.75% and > 1.25% coverage (as previously applied), and not applying any coverage filter. The overlap of the LOH blocks was taken as proxy for the reproducibility of the method. This overlap was determined with the jaccard metric of bedtools v2.25.0 (Quinlan & Hall, 2010). Our analysis indicated that the results were only reproducible removing the coverage filter (see Results section). Thus, for LOH blocks definition, bedtools merge v2.25.0 (Quinlan & Hall, 2010) with a window of 100 bp was used to define heterozygous regions, and by opposite, LOH blocks would be all non-heterozygous regions in the genome. The minimum LOH block size was established at different thresholds, namely, 100 bp, 150 bp, 200 bp, 500 bp, 1 kb and 5 kb, and therefore 6 parallel analyses were performed. The “degree of homozygosity” per gene was defined as

the number of positions in a given gene coinciding with LOH blocks per gene size, and it was calculated with bedtools coverage v2.25.0 (Quinlan & Hall, 2010). The overlap of homozygous and heterozygous genes between strains or clades was performed with hypergeometric tests using phyper function of R, setting lower.tail to false.

Functional gene annotation

GO terms for *C. orthopsilosis* were retrieved from Candida Genome Database (http://www.candidagenome.org/) on September 27th, 2019. To complement this information, functional annotation was performed for both species (*C. orthopsilosis* and *C. metapsilosis*) with InterProScan v5.21.6 (Jones et al., 2014), and EggNOG-mapper v1.0.3 using Diamond algorithm (Huerta-Cepas et al., 2017). From this analysis we retrieved information not only on GO terms, but also on KEGG pathways, Pfam domains and orthologs between the two species.

Enrichment analysis

To inspect if any function was preferentially undergoing LOH in the hybrid strains, for each hybridization event the average and the median of homozygosity per gene was calculated. Enrichment analysis was performed comparing two different datasets, namely, heterozygous and homozygous genes using FatiGO (Al-Shahrour et al., 2007) in order to find GO terms, KEGG terms, or Pfam domains over-expressed in one of the lists. However, to do that it was necessary to determine a threshold between homozygous and heterozygous genes. Thus, this threshold was set at 100%, 90%, 80% and 50% of mean or median homozygosity for each gene. Moreover, as conserved genes could be influencing the results, we have decided to exclude from the analysis all the genes without non-synonymous SNPs. Analyses were performed considering only genes with at least one non-synonymous mutation in at least one strain, and later considering all the genes. As we were also interested in assessing if any function was kept as heterozygous, a similar approach was performed for 0%, 10% and 20% "gene homozygosity".

Pentatricopeptide repeat proteins

Due to their possible role in genomic incompatibilities, PPR proteins were a special focus of this work. PPR proteins were selected considering the list previously described for *S. cerevisiae* and *C. albicans* (Lipinski et al., 2011). The orthologs of these genes were retrieved from the analysis with EggNOG-mapper v1.0.3 (Huerta-Cepas et al., 2017). A search for PPR domains in the remaining proteome was performed with HMMER (http://hmmer.org). Proteins predicted by functional gene annotation analysis as PPRs were also added to this dataset. In total we analyzed 13 PPR proteins. The level of homozygosity was inspected for each gene in each strain. In case of genes 100% homozygous in all strains from a given clade, the respective genomic region was inspected with IGV version 2.0.30 (Thorvaldsdottir et al., 2013).

Determination of the different hybridization events

Previous authors have identified one hybridization event in the origin of *C. metapsilosis* and at least four independent hybridization events in the origin of *C. orthopsilosis* (Schröder et al., 2016; Pryszcz et al., 2015). To check whether the new strains considered in this project belong to one of these hybridization events or to new ones, bedtools jaccard v2.25.0 (Quinlan & Hall, 2010) was used to quantify the number of nucleotide positions in the union of LOH blocks in strain pairwise comparisons.

Furthermore, for each *C. orthopsilosis* and *C. metapsilosis* strain, the homozygous variants were substituted in the respective reference genome with the FastaAlternateReferenceMaker tool of GATK v3.6 (McKenna et al., 2010). Positions with heterozygous variants or INDELs in at least one of the respective strains were removed from the final alignments with bedtools subtract v2.25.0 (Quinlan & Hall, 2010). In the end, two concatenated alignments of 12,028,303 and 12,336,395 nucleotides were obtained for *C. orthopsilosis* and *C. metapsilosis* strains, respectively. A Maximum-likelihood tree representative of each alignment was generated with RAxML v8.2.8 software (Stamatakis, 2014) using GTRCAT model.

Analysis of ITS sequences of environmental isolates

Publicly available ITS sequences for environmental isolates were retrieved from NCBI database (LC415306.1, AB863470.1, AB863471.1, KJ194280.1, KJ194328.1, KJ194334.1, and KJ194336.1 (Kaewkrajay et al., 2020; Khunnamwong et al., 2018; Xu et al., 2014). After an online BLASTn of each of these sequences, we could separate them in two groups, as some corresponded to the 26S sequence (LC415306.1, AB863470.1, and AB863471.1) and the others to the ITS1 (KJ194280.1, KJ194328.1, KJ194334.1, and KJ194336.1). The ITS1 and 26S regions were retrieved for each *C. metapsilosis* strain after replacing the respective homozygous variants in the reference genome with FastaAlternateReferenceMaker tool of GATK v3.6 (McKenna et al., 2010). Alignments were performed with MAFFT online interface (Katoh et al., 2019; Katoh & Standley 2013). Positions with gaps were removed with trimAL v1.4.rev15 (Capella-Gutiérrez et al., 2009). A phylogenetic tree representative of each alignment was generated with RAxML v8.2.8 software (Stamatakis, 2014) using GTRCAT model. Of note, as in the ITS group of strains there was an apparent long-branch attraction, the analysis was performed independently for each public sequence.

Estimation of the divergence time for each hybridization event

To understand if the LOH level is directly correlated to the age of the different hybridization events, we compared the shared LOH blocks of all strains from a given clade and counted for each strain the number of SNPs which are not shared between all of them, assuming that in this case they represent mutations acquired after the divergence. To estimate the divergence time for each clade the average number of new SNPs per kilo-base per strain was considered. The estimated divergence time for each clade was compared with the average amount of LOH acquired after the clade divergence.

Detection of chromosome recombination in heterozygous positions

To detect events of recombination within heterozygous blocks, we used an *in-house* script. Briefly, for each strain and each heterozygous block with more than 500bp, we phased all the 0/1 heterozygous positions using HapCUT2 (release April 4th, 2019, (Edge et al., 2017)). The average allele frequency of each block was calculated by dividing the number of reads supporting each haplotype in the heterozygous regions. All the blocks with ploidy different from 2 were excluded (allele frequency < 40% or allele frequency > 60%). We considered as potential regions of recombination, all the phased blocks with more than 10 consecutive SNPs supporting each of the possible phased genotypes (0|1 and 1|0). All the obtained results were inspected with IGV version 2.0.30 (Thorvaldsdottir et al., 2013).

C. orthopsilosis sporulation assays

In order to investigate the potential sexual reproduction of *C. orthopsilosis*, we decided to induce meiosis in MCO456, s424 and s425 strains. A diploid *S. cerevisiae* was used as control. All strains were taken from our glycerol collection and grown on YPD agar plates for 2 days at 30°C. Further, the samples were streak onto potassium acetate agar plates (yeast extract [0.25%], glucose [0.1%], potassium acetate [10%], agar [2%]) (Guitard et al., 2015) and incubated at 30°C for 4 weeks. Each week, a thin smear of the culture was stained using the Schaeffer and Fulton Spore Stain Kit (Sigma-Aldrich, cat. No 04551) following the producers' protocol. Green/blue color indicating ascospores was seen in the positive control whereas no spores were observed in the investigated strains.

Data availability

The *de novo* genome assembly and raw sequencing reads generated for this study can be found in SRA under the BioProject PRJNA520893.

10 HaploTypo: a variant-calling pipeline for phased genomes

Pegueroles, C.*, Mixão, V.*, Carreté, L.*, Molina, M., & Gabaldón, T. (2020). HaploTypo: a variant-calling pipeline for phased genomes. *Bioinformatics*, *36(8)*, 2569–2571. doi: 10.1093/bioinformatics/btz933 *Equal contribution

10 HaploTypo: a variant-calling pipeline for phased genomes

10.1 Abstract

Summary

An increasing number of phased (i.e. with resolved haplotypes) reference genomes are available. However, the most genetic variant calling tools do not explicitly account for haplotype structure. Here, we present HaploTypo, a pipeline tailored to resolve haplotypes in genetic variation analyses. HaploTypo infers the haplotype correspondence for each heterozygous variant called on a phased reference genome.

Availability and implementation

HaploTypo is implemented in Python 2.7 and Python 3.5, and is freely available at https://github.com/gabaldonlab/haplotypo, and as a Docker image.

Supplementary information

Supplementary data are available at *Bioinformatics* online.

10.2 Motivation

The heterozygosity (i.e. the presence of alternative alleles at the same locus) present in diploid organisms can complicate genome analyses, particularly when the levels of heterozygosity are high. Over the last years, several bioinformatics tools have been developed to account for this sequence complexity. These include pipelines and algorithms to assist during the genome assembly process (Pryszcz & Gabaldón, 2016; Safonova et al., 2015), subsequent phasing of assembled genomes (Chin et al., 2016; Edge et al., 2017; Pan et al., 2014) or allele-specific transcriptomic analysis (Deonovic et al., 2017; Romanel et al., 2015). However, and to the best of our

knowledge, available variant calling tools do not explicitly account for phased genomes. As a result, the user has to decide between using the combined phased haplotypes as reference and thereby losing heterozygosity information or, alternatively, using only one of the haplotypes as reference and sacrificing haplotype information. An illustrating example of such problem is studies on the heterozygous yeast pathogen *Candida albicans.* Although the diploid genome of this pathogen was phased in 2013 (Muzzey et al., 2013), subsequent studies have only used one of the haplotypes (Bensasson et al., 2019; Ropars et al., 2018), thereby losing the valuable haplotype information. Given the increasing amount of highly heterozygous genomes, including those from hybrids (Mixão & Gabaldón, 2018), and the relevance of phased information to reconstruct their population structures and evolutionary histories, there is an urgent need for solutions that allow the exploitation of phased genomes in genomic variation analysis. To fill in this gap, we developed HaploTypo, a python-based pipeline that, in the presence of a phased reference genome, provides detailed genome variation resolved at the haplotype level. HaploTypo is not a *de novo* genome phasing tool, but a tool to phase variants in re-sequencing analysis, using information of an already phased genome, resulting in a fast and accurate assessment of heterozygosity levels and reconstruction of haplotypes.

10.3 Implementation

HaploTypo requires as input the phased haplotypes of a diploid genome, and filtered genomic paired-end sequencing reads or, alternatively, their alignment to each of the reference haplotypes. The pipeline is divided in four modules, which can be run in block or separately (Figure 10_1). The first module aligns the genomic paired-end reads independently to each of the phased haplotypes using BWA-MEM (Li, 2013). The second module performs variant calling on the two generated alignments using GATK (McKenna et al., 2010), BCFtools (Li, 2011) or FreeBayes (Garrison & Marth, 2012) followed by variant filtration. From here, variability information is obtained for each reference haplotype independently. The third module of HaploTypo implements a variant phasing algorithm that, based on the comparison of reference haplotypes and

previously called variants, infers which variants correspond to each haplotype. Phased (and unphased if required) genotypes from the two phased haplotypes are provided as independent VCF files. Additionally, unphased and unsolved positions for each haplotype are reported as bed files. A final module uses the VCF files generated in module 3 to reconstruct the haplotypes and provide them in fasta format. Detailed information on HaploTypo implementation is available in the pipeline's manual.

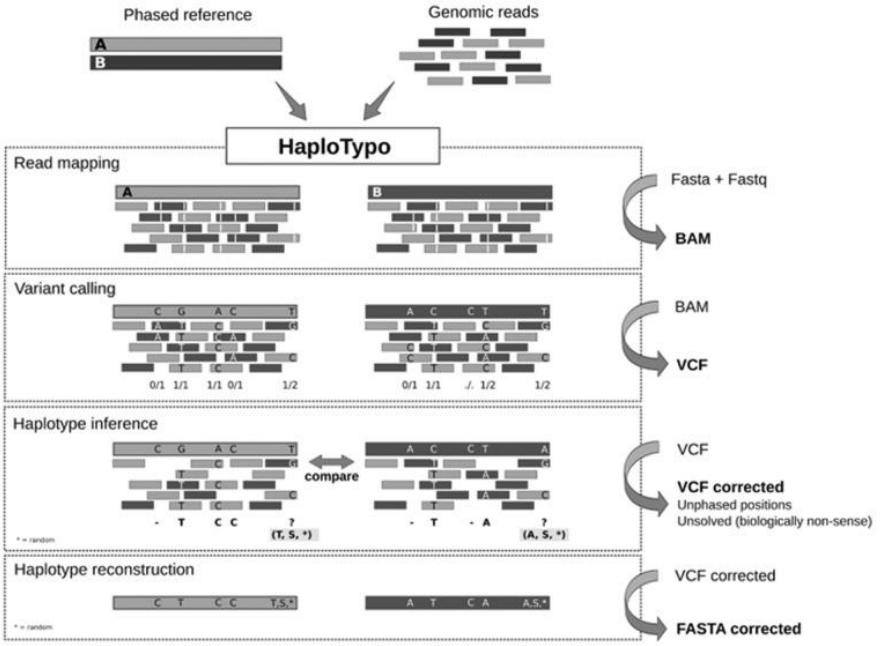

Figure 10_1. Schematic representation of the four modules of HaploTypo pipeline. The steps are described in the main text and in the pipeline's manual.

10.4 Validation and Results

We validated HaploTypo using simulated phased genome sequences with known variable positions. To explore the influence of divergence between the two phased haplotypes in the downstream analysis, we simulated diploid reference genomes with haplotypes diverging 0.5, 1 or 5% at the nucleotide level. These simulated

phased reference genomes were derived with fasta2diverged.py [https://github.com/lpryszcz/bin, (Pryszcz et al., 2015)] from *C. albicans* haplotype A (Muzzey et al., 2013). The same script was used to simulate diploid strains, which differed from the respective reference genome in 1 position per kilo-base, referred to as the 'simple' dataset. Given that, for these simulated strains, most of the polymorphisms are of the type 0/1 (where 0 is reference allele and 1 is alternative allele), we also simulated divergent strains where the polymorphisms between the two haplotypes and the reference could be 0/1 (60% of the total variation), 1/1 (38% of the total variation) or 1/2 (2% of the total variation, where 2 is an alternative allele different from 1), referred to as the 'complex' dataset. The relative proportions of the different variant types were based on real data from *C. albicans* sequenced strains (Ropars et al., 2018). We next simulated sequencing reads using wgsim v0.3.1-r13 (https://github.com/lh3/wgsim). By using these simulated references and reads, we compared the performance of the HaploTypo pipeline to: (i) mapping libraries to only one of the haplotypes (standard procedure) and, (ii) mapping the libraries to the phased genome reference, which combines the two haplotypes (alternative approach). In all cases, we assessed the performance with the three mentioned variant callers.

As expected, when using a haploid reference, sensitivity varied between 96.37 and 99.38%, depending on the variant caller and the divergence between haplotypes (Supplementary Table S1). However, as discussed above, this approach serves to assess heterozygosity levels and the location of heterozygous SNPs, but this information is unphased. When using a diploid reference, the divergence between the two haplotypes highly influenced the outcome, with better results being achieved at higher nucleotide divergences, with sensitivity ranging from 6.5 to 19.3% for 0.5% divergence, from 36.7 to 66.4% for 1% divergence and from 98.3 to 98.8% for 5% divergence (Supplementary Table S1). GATK had the poorest results, specially at low divergence levels (Supplementary Table S1). It is worth noting that accuracy and specificity remained high and stable (>99% and 100% respectively) independently of the levels of divergence and the variant caller used. When using HaploTypo, reads are mapped independently to the two haploid references of a phased genome (approach with the best results, as shown above) and outputs the two haplotypes, correctly phasing

>99% of the positions independently of the variant caller used, with few exceptions (Supplementary Table S2). The unphased cases always represented ambiguous situations that cannot possibly be resolved with this type of data (see manual for details), and the user can decide whether to include them in the VCF or not. In addition, HaploTypo also reports positions that have incoherent results in the two haplotypes and therefore are likely to be mapping or variant calling errors (unsolved positions, see Table 1 from the manual for details). HaploTypo benchmarking was performed on a workstation [Intel(R) Xeon(R) CPU E5-1650 v3] and 64 GB of RAM with default number of threads. The total running time ranged from 1 to 15 h, depending on the level of heterozygosity of the dataset and the variant caller (Supplementary Table S3). Hence HaploTypo is a user-friendly tool that eases variant analyses and allows to incorporate haplotype-specific information when a phased reference genome is available.

Part II

Discussion

11 Summarizing discussion

Species of the genus *Candida* are important for human health, not only because some of them are an integral part of our microbiota, but also because under certain circumstances, such as weakening of the host immune system, they can become opportunistic pathogens and cause life-threatening infections (Consortium OPATHY & Gabaldón, 2019; Turner & Butler, 2014). In the last years, some of these species have represented an important burden to hospitals, being a threat for diverse types of patients and causing deadly outbreaks (Consortium OPATHY & Gabaldón, 2019; Pfaller et al., 2010b; Turner & Butler, 2014). Besides well-known pathogens, new pathogenic lineages are continuously emerging, posing new threats and leading to additional challenges for diagnosis and treatment (Pfaller et al., 2010b). Related to this, hybridization has been pointed as one of the possible routes through which virulence towards humans may emerge (Mixão & Gabaldón, 2018; Pryszcz et al., 2015). The diversification of the possible causes of *Candida* infections urges the need for a better knowledge of their genomic and phenotypic features. In this context, it is not only important to study *Candida* pathogens, but also their non-pathogenic relatives, as this is crucial to understand their evolution. This thesis aimed to study the genomic features of *Candida* pathogens in light of hybridization. Two main questions were addressed: i) how widespread are hybrids among *Candida* spp., and ii) what processes drive the evolution of their genomes. Overall, this thesis used comparative genomics techniques to explore genomes of thirteen diverse *Candida* species.

Hybridization goes beyond inter-species mating

The classification of a lineage as hybrid has important implications on how to perform biological research on it. Hybridization is an event of non-vertical evolution in which two genomes are brought together, originating a highly heterozygous organism. This heterozygosity may negatively impact hybrid's fitness but, at the same time, it might be an important source of genomic plasticity and adaptation to new environments (Mixão & Gabaldón, 2018). Thus, from the moment a

lineage is identified as a result of hybridization, its evolutionary path must be studied from a completely different perspective.

Hybridization is often described as the cross between two different species (inter-species hybridization) (Mallet, 2007; Stukenbrock, 2016). Nevertheless, the biological significance of hybridization results from the coexistence of different genomic material, regardless of the categorization as different species of the donor organisms. Even if the degree of divergence between the two parental lineages may indeed influence the extend of the aftermath of hybridization, it cannot be used to define the concept of hybridization *per se*. Therefore, the concept of "species", which although relatively easy to apply in animals and plants, is particularly difficult to apply in fungi, should not influence the concept of hybridization. For instance, this thesis has shown evidence for the occurrence of mating between *Candida* lineages, whose divergence at the nucleotide level is within the range of the high inter-strain variability described for other yeasts (Liti et al., 2006). This did not allow the conclusion that they correspond to inter-species hybrids. However, application of a broader hybridization concept that does not rely on the definition of species has allowed a better understanding of the evolution of *Candida* pathogens. Considering this, the definition of hybridization should be broadened again to what was considered in the past by Gregor Mendel in its first genetic studies, i.e. the mating of two organisms of different breeds, varieties, species or genera, because what is really important is the fact that two genomes that diverged for some time were combined in a new cell (Mendel, 1866; Mixão et al., 2019; Pryszcz et al., 2015; Schröder et al., 2016).

Hybrids are widespread among *Candida* spp.

As mentioned above, thirteen *Candida* species were analyzed in this thesis. To this end, both previously available sequencing data and newly obtained sequences were used, which allowed the largest exploration thus far of the extent of hybridization among *Candida* species. Importantly, from these species, eight were found to be hybrids, and six of them were described as hybrids for the first time in this thesis. The first hybrids ever reported in *Candida* spp. correspond to clinical isolates of *C. orthopsilosis* and *C. metapsilosis* (Pryszcz et al., 2014, 2015; Schröder et al., 2016). At the beginning

of this project, four hybridization events were described for *C. orthopsilosis*, while a unique hybridization event was proposed for *C. metapsilosis* (Pryszcz et al., 2015; Schröder et al., 2016). By then, these were the only hybrids described in the clade. With the sequencing of additional samples from these species, we revealed the existence of a differentiated clade in *C. metapsilosis*, but with the data at hand we consider unlikely that it results from a second hybridization event. However, among the 18 newly sequenced strains we were not able to identify any strain belonging to any of the missing parents of these hybrids. This reinforces the previously proposed idea that the unknown parental lineages of these hybrids might correspond to non-pathogenic strains, which would explain their absence from clinical isolates (Pryszcz et al., 2015). In favor of this hypothesis is the isolation of environmental strains of *C. metapsilosis*. In this regard, we sequenced one strain isolated from a marine environment and proved that it is also a hybrid descending from the same hybridization event as *C. metapsilosis* clinical isolates. Although we cannot completely exclude the possibility that this hybrid is the result of contamination and not a real environmental isolate, the fact that this niche is consistent with other studies reporting environmental *C. metapsilosis* isolations (Kaewkrajay et al., 2020; Khunnamwong et al., 2018) makes this possibility less likely. It is unknown whether other environmental *C. metapsilosis* isolates are also hybrids, but the finding of a hybrid strain in the environment supports the previously proposed hypothesis of environmental hybridization events as a possible source of emerging hybrid pathogens (Pryszcz et al., 2015).

To broaden our search for possible parental lineages, we sequenced four closely related species, which did not correspond to clinical isolates, namely *C. oxycetoniae*, *C. jiufengensis*, *C. pseudojiufengensis* and *C. theae*. Contrary to our expectations, none of these species corresponded to any of the unknown parental lineages. But, surprisingly, the genomic patterns of *C. theae* indicated that this species was also a hybrid. This species is phylogenetically very close to both *C. orthopsilosis* and *C. metapsilosis* (Figure 1_1). Given that *C. theae* was isolated from a common beverage and it belongs to a clade known for harboring pathogenic species (Chang et al., 2012), some concern should be raised regarding the ability of this hybrid to impact human health. Indeed, even if the main habitat of a species is in the environment, this does not exclude the chance that

this species may cause disease. A good example is *C. subhashii*, another member of the CUG-Ser clade that is not particularly associated to humans but was responsible for a case of peritonitis (Adam et al., 2009). In this thesis we sequenced and analyzed this pathogenic isolate, which we also found to be a hybrid. These findings reinforced the idea that lineages of this clade are prone to hybridize, and for a yet unknown reason, hybrids are more associated to human and human-related environments than the parental lineages. To understand this possible association, it is crucial to increase our capacity to isolate and sequence new species.

We also sequenced three rare yeast pathogens: *D. rugosa*, *T. ciferrii* and *C. inconspicua.* Our results revealed that one of them, *C. inconspicua*, is a hybrid with two clades which possibly correspond to two different hybridization events. The analysis of a putative isolate from Canada revealed a third hybrid clade, which only shares one of the parental lineages with *C. inconspicua*. Additional analyses of other members of the species complex where *C. inconspicua* belongs revealed a much more complex picture in which all sequenced samples of the complex correspond to hybrid strains. While *P. norvegensis* is a hybrid that has undergone massive LOH (and therefore the hybridization tracks were almost completely erased from the genome), *P. cactophila* is highly heterozygous, and our results suggest that it is a "double hybrid" resulting from the cross of a *C. inconspicua* hybrid and another unknown lineage. Interestingly, although one of the parents was a hybrid, *P. cactophila* has an allele frequency compatible with the mating of two homozygous lineages. Furthermore, *P. cactophila* presented a recombination in the *MAT* locus of the sub-genome from the hybrid parental. This recombination event was similar to a recombination in the same locus described in the past (and also confirmed in this thesis project) for all *C. metapsilosis* hybrids (Pryszcz et al., 2015). Disruption of the *MAT* locus has previously been associated to restoration of fertility in hybrids (Ortiz-Merino et al., 2017). Thus, we hypothesize that the recombination event in the *MAT* locus may have enabled *C. metapsilosis* and *C. inconspicua* to restore their mating ability. In the case of *C. inconspicua*, this event possibly helped to restore a homozygous state, and subsequently mate with a different homozygous lineage, originating *P. cactophila*.

The genome of *C. albicans*, the main *Candida* pathogen, is known for having blocks of heterozygosity separated by LOH (Ropars et al., 2018), which is a characteristic of hybrid genomes. Considering the high number of hybrids that our analyses revealed, we decided to assess whether the heterozygosity of *C. albicans* could also be a result of a hybridization event. Taking advantage of a large amount of sequencing data available in public databases, we compared the genome of several *C. albicans* isolates, including clinical and environmental samples. Our results showed compelling evidence for a hybrid ancestor of *C. albicans*, which has been evolving for relatively long time when compared to other hybrids of the CUG-Ser clade. A previous study has shown that the closely related lineage *C. africana*, which is sometimes considered the same species as *C. albicans*, also presents some heterozygosity but the levels of LOH are much higher than those observed in *C. albicans* (Ropars et al., 2018). Therefore, we hypothesized that *C. africana* could be one of the parental lineages. Our results showed that *C. africana* is a mosaic of *C. albicans* haplotypes, and therefore it must share the same hybrid ancestor as this species but, similarly to *P. norvegensis*, it underwent a massive LOH. It is important to note that recombination has been described in *C. albicans* (Anderson et al., 2019), and the current haplotypes might be a mixture of the ancestral ones. If this is the case, we cannot be certain that the mosaic observed in *C. africana* corresponds indeed to multiple events of LOH towards different parentals, or to the result of multiple recombinations between the haplotypes with which we were comparing. Independently of the role of *C. africana* in the evolutionary path of *C. albicans*, these results suggest that a hybridization event was in the origin of the most important *Candida* pathogen. The absence of known parental lineages can be once again an indicator that, similarly to other hybrids of the CUG-Ser clade, these are non-pathogenic or less virulent than the hybrid, and the hybridization event was the trigger that made *C. albicans* such a successful opportunistic pathogen.

Altogether, these results show that multiple hybridization events have occurred in *Candida* spp.. These events gave rise to highly heterozygous lineages which often correspond to opportunistic pathogens. The proportion of hybrids and non-hybrid clinical isolates suggests that these hybrids might be more virulent or tend to be more associated to humans than their non-hybrid parentals. For instance, besides *C. inconspicua, C. metapsilosis* and *C. albicans* in which all

strains are hybrids, in *C. orthopsilosis* where some homozygous clinical isolates were found, they only corresponded to approximately 12% of the cases. This apparent tendency to isolate hybrids in the clinic raises the question of whether highly heterozygous lineages have higher chances of success when compared to the homozygous ones. Furthermore, the high number of hybrids identified suggests that *Candida* species are prone to hybridize.

Evolution of hybrid genomes of the CUG-Ser clade

Usually, the high heterozygosity levels in hybrid genomes are accompanied by genomic incompatibilities that may involve fitness costs (Mixão & Gabaldón, 2018). This idea is at odds with the high number of hybrids identified in *Candida* species and their widespread nature. The identification of multiple events in the *C. parapsilosis* species complex represented an unprecedented opportunity to apply comparative genomics to assess convergent trends in independently formed hybrid clades. That is, if homozygosity or heterozygosity of certain genes was advantageous, then one would expect these being consistently selected in independently formed clades. Our results revealed that LOH patterns are mostly shaped by genetic drift. This either suggests that the fitness costs of epistatic interactions in heterozygous regions are low, or that selection cannot effectively select beneficial LOH events, or a combination of both. In addition, the strength of selection might be low, preventing the efficient selection against existing incompatibilities. However, this last scenario could negatively impact hybrid fitness.

Several parameters determine how efficiently selection can eliminate deleterious variants and select beneficial ones. For instance, low effective population sizes coupled to numerous bottlenecks, in combination to the absence of sexual recombination would be factors that limit the strength of selection and are likely to be present in these hybrids. Moreover, the nature of the mutational process can also determine the balance between selection and drift. Selection acts on genotypes and not on individual mutations. If LOH events affect many genes simultaneously, then selection for one beneficial allele will carry all other alleles affected by the same LOH, making it difficult to infer the target of selection from a limited set of patterns.

If multiple LOH events appear in a short period of time, this effect would be exacerbated. Related to this, our results have shown that LOH blocks in these hybrids are not linearly accumulated through time and have patterns compatible to the occurrence of chromosomal recombination at very high rates. It was not possible to identify the underlying mechanism associated to these possible recombination events. Nevertheless, we suggest that a process similar to RTG, in which the cell enters the meiotic cycle but exits from it before sporulation (Simchen, 2009), is consistent with our observations of high recombination rates in the absence of sporulation. In such a case, these hybrids could acquire multiple LOH regions in a single step. Therefore, we think that all the above-mentioned factors may have played a role in the difficulty of detecting convergent LOH patterns. If, indeed, it is the overlap of LOH with multiple genes that influenced our results, a larger sample size could help solving this issue. Otherwise, the absence of strong incompatibilities is the best hypothesis.

As mentioned before, similarly to previous studies, we were not able to induce sporulation in hybrids of *C. orthopsilosis* (Butler et al., 2009; Pryszcz et al., 2014). This lack of evidence for the occurrence of meiosis in species of the CUG-Ser clade raises the question of how the hybrid lineages originated. *C. albicans* and two closely related species are known for having an alternative sexual cycle (parasexuality followed by parameiosis) in which two diploid cells mate forming a tetraploid that restores the diploid status by concerted chromosome loss (Anderson et al., 2019; Berman & Hadany, 2012; Seervai et al., 2013). A mechanism similar to this one could be proposed as the source of hybridization events in the CUG-Ser clade. However, as discussed for the origin of *C. albicans* in Chapter 8, the most part of the analyzed hybrids have the two haplotypes for all chromosomes, which would imply that the chromosome loss would always select a copy of each haplotype. As we consider this hypothesis very unlikely, we suggest that these hybrids may be the result of mating between haploid lineages. This last hypothesis can be refuted by the fact that these species and their close relatives tend to be diploid. However, competent *C. albicans* haploids have been induced in laboratory (Hickman et al., 2013). If this hypothesis of mating of haploid cells is real, this means that we are missing a big portion of all the picture that is the CUG-Ser clade. This could be related to our tendency to study pathogens instead of environmental

isolates, or to gaps in our capacity to isolate and maintain such lineages in the laboratory. The clarification of the process underlying hybridization events in the CUG-Ser clade would not only help to understand the evolution of hybrids and, particularly, hybrid pathogens, but also the evolution of *Candida* spp..

Study the past to understand the future

Advances in NGS have allowed us to identify hybrids which otherwise would not be possible to identify. This thesis has shown that the number of hybrid lineages is higher than what was previously anticipated. Sequencing tends to be biased towards clinical isolates, but in most cases those isolates are sequenced before knowing their hybrid nature, so one could argue that sequences in the databases are not expected to be enriched in hybrids. Interestingly, from the 50 *Candida* genomes publicly available or sequenced in this project, so far, six (12%) - all of them analyzed in this thesis - have been described to be hybrids, and this number is expected to be even larger. The high plasticity characteristic of hybrid genomes makes it important to distinguish hybrids from non-hybrid pathogens, as their response to stress conditions is hypothesized to be more flexible (Mixão & Gabaldón, 2018). It is still unknown whether this genomic plasticity is influencing the outcome of the disease, but it is worth noting that highly heterozygous and/or polyploid strains appear to be associated to more hostile environments (Diezmann & Dietrich, 2009; Forche et al., 2011; Zhu et al., 2016), and hybridization is one of the mechanisms through which high levels of heterozygosity can be quickly acquired (Mixão & Gabaldón, 2018).

An essential step to understand hybrids and be prepared for potential future outbreaks of hybrid pathogens, is to explore their evolution. In order to do so, it is important to fill in the gap related to the absence of known parental lineages. This would help to understand the genomic characteristics of the parentals and assess the changes induced by hybridization. In this context, it would be of extreme relevance to sequence as many species as possible. Furthermore, the development of new techniques to isolate and culture yeasts could open doors to explore a still unknown world of diversity. While the parental lineages are not sequenced, alternative Bioinformatics tools can help to reconstruct the ancestral haplotypes of each hybrid using

genome phasing techniques. A challenge in this field is still the presence of blocks of LOH, which do not allow to connect the two phased haplotypes flanking those blocks, if the block size exceeds the size of the pair of short reads. This problem can be partially surpassed by using long-read sequencing technology. However, the presence of high error rates might influence the outcome. Another challenge associated to genome phasing for the reconstruction of ancestral haplotypes, is the presence of haplotype recombination, as mentioned before for *C. albicans.* Even with all these challenges, the number of phased genomes is increasing. As an example, the genome of *C. albicans* was phased already seven years ago (Muzzey et al., 2013). During all this time even if the genome was phased the scientific community has only been working with one of the haplotypes because tools are not able to deal with phased reference genomes (Bensasson et al., 2019; Ropars et al., 2018). To fill in this gap, in this thesis we developed HaploTypo (Pegueroles et al., 2020), which is able to assess genome heterozygosity without losing phasing information. This tool will be of extreme relevance for future studies in hybrid genomes.

The connection between hybrids and pathogenicity is not clear, but pathogenicity has emerged multiple times independently (Gabaldón et al., 2013). Previous studies suggest that hybridization can lead to the emergence of new pathogens (Mixão & Gabaldón, 2018; Pryszcz et al., 2015), and indeed the finding of many hybrid pathogenic lineages among non-hybrid and non-pathogenic strains supports this idea. In a world where climate changes and globalization are a reality, the movement of lineages to places where they have never been is a constant (King et al., 2015). This might facilitate the contact between diverged lineages, which can hypothetically hybridize, and generate new organisms which can adapt to new environments, like the human body. Therefore, some concern should be raised regarding the emergence of new pathogens, and efforts should be made to explore the association between hybridization and the emergence of pathogenicity.

12 Conclusions

The main conclusions of this thesis are:

- Hybridization is widespread in *Candida* spp. and includes clades relevant for human health.

- *P. cactophila* species complex has at least five distinct lineages, all of them related to hybridization events.

- The emerging pathogen *C. inconspicua* has at least three different lineages that correspond to different hybridization events and only comprise pathogenic clinical isolates.

- *C. theae* is a hybrid member of *C. parapsilosis* species complex, similarly to *C. orthopsilosis* and *C. metapsilosis.*

- A hybridization event is at the origin of *C. albicans* lineage.

- Mild genomic incompatibilities between the parental genomes might be associated to the success of hybrids of the CUG-Ser clade.

- Loss of heterozygosity in *Candida* hybrids is possibly not linearly accumulated through mitotic recombination, but rather through a mechanism similar to “return to growth” or parameiosis.

Appendices

Appendix 1

The supplementary material supporting the results of this thesis project is available at:

https://www.dropbox.com/sh/sytc01ck0pfa4qj/AABn4_36N8778EADEzR5V_una?dl=0

Appendix 2

List of scientific publications associated to the work developed during this thesis project:

1. **Mixão, V.** & Gabaldón T. (2018). Hybridization and emergence of virulence in opportunistic human yeast pathogens. *Yeast*, *35(1),* 5-20. doi: 10.1002/yea.3242

2. Stavrou, A.A.*, **Mixão, V.*,** Boekhout, T., & Gabaldón, T. (2018). Misidentification of genome assemblies in public databases: The case of *Naumovozyma dairenensis* and proposal of a protocol to correct misidentifications. *Yeast*, *35(6),* 1-5. doi: 10.1002/yea.3303 *Equal contribution

3. **Mixão, V.,** Hansen, A. P., Saus, E., Boekhout, T., Lass-Flörl, C., & Gabaldón, T. (2019). Whole-Genome Sequencing of the Opportunistic Yeast Pathogen *Candida inconspicua* Uncovers Its Hybrid Origin. *Frontiers in Genetics*, *10*, 383. doi: 10.3389/fgene.2019.00383

4. **Consortium OPATHY***, & Gabaldón, T. (2019). Recent trends in molecular diagnostics of yeast infections: from PCR to NGS. *FEMS Microbiology Reviews*, *43(5),* 517–547. *Verónica Mixão is part of OPATHY consortium

5. **Mixão, V.,** Saus, E., Hansen, A. P., Lass-Flörl, C., & Gabaldón, T. (2019). Genome Assemblies of Two Rare Opportunistic Yeast Pathogens: *Diutina rugosa* (syn. *Candida rugosa*) and *Trichomonascus ciferrii* (syn. *Candida ciferrii*). *G3: Genes|Genomes|Genetics*, *9*(12), 3921–3927. doi: 10.1534/g3.119.400762

6. Pegueroles, C.*, **Mixão, V.*,** Carreté, L.*, Molina, M., & Gabaldón, T. (2020). HaploTypo: a variant-calling pipeline for phased genomes. *Bioinformatics*, *36*(8), 2569–2571. doi: 10.1093/bioinformatics/btz933 *Equal contribution

7. **Mixão, V.,** & Gabaldón, T. (2020). Genomic evidence for a hybrid origin of the yeast opportunistic pathogen *Candida albicans*. *BMC Biology*, *18*, 48. doi: 10.1186/s12915-020-00776-6

8. Priest, S.J., Coelho, M.A., **Mixão, V.**, Clancey, S., Xu, Y., Sun, S., Gabaldón, T., & Heitman, J. (2020). Factors enforcing the species boundary between the human pathogens *Cryptococcus neoformans* and *Cryptococcus deneoformans*. *bioRxiv*, doi: 10.1101/2020.05.21.108084

9. Muñoz-Barrios, A., Sopeña-Torres, S., Ramos, B., López, G., del Hierro, I., Díaz-González, S., González-Melendi, P., Mélida, H., Fernández-Calleja, V., **Mixão, V.**, Martín-Dacal, M., Marcet-Houben, M., Gabaldón, T., Sacristán, S., & Molina, A. (2020) Differential expression of fungal genes determines the lifestyle of *Plectosphaerella* strains during *Arabidopsis thaliana* colonization. *(Under revision at Molecular Plant-Microbe Interactions)*

10. **Mixão, V.,** Ksiezopolska, E., Saus, E., Boekhout, T., Gacser, A., & Gabaldón, T. Effect of drift, selection, and recombination on the evolution of genomes of hybrid yeast pathogens. *(In preparation)*

11. **Mixão, V.,** Boekhout, T., & Gabaldón, T. Whole-genome analysis reveals that hybridization is widespread among *Candida* species. *(In preparation)*

12. Pérez-Hansen, A., **Mixão, V.,** Sarttori, B., Gabaldón, T., Lass-Flor, C., & Lackner, M. Low variation in the sequences of the hotspot regions of *FKS1* in clinical isolates of two rare *Candida* species: *Candida inconspicua* and *Candida rugosa*. *(In preparation)*

References

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C. A., Buggs, R., Butlin, R. K., Dieckmann, U., Eroukhmanoff, F., Grill, A., Cahan, S. H., Hermansen, J. S., Hewitt, G., Hudson, A. G., Jiggins, C., Jones, J., Keller, B., Marczewski, T., Mallet, J. Martinez-Rodriguez, P., Möst, M., Mullen, S., Nichols, R., Nolte, A. W., Parisod, C., Pfennig, K., Rice, A. M., Ritchie, M. G., Seifert, B., Smadja, C. M., Stelkens, R., Szymura, J. M., Väinölä, R., Wolf, J. B. W., & Zinner, D. (2013). Hybridization and speciation. *Journal of Evolutionary Biology, 26(2),* 229–246. 10.1111/j.1420-9101.2012.02599.x

Adam, H., Groenewald, M., Mohan, S., Richardson, S., Bunn, U., Gibas, C. F. C., Poutanen, S., & Sigler, L. (2009). Identification of a new species, *Candida subhashii*, as a cause of peritonitis. *Medical Mycology: Official Publication of the International Society for Human and Animal Mycology, 47(3),* 305–311. 10.1080/13693780802380545

Aguilera, A., & Rothstein, R. (2007). Molecular Genetics of Recombination. Springer Science & Business Media.

Al-Shahrour, F., Diaz-Uriarte, R., & Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics, 20,* 578–580. 10.1093/bioinformatics/btg455

Al-Yasiri, M. H., Normand, A.-C., L'Ollivier, C., Lachaud, L., Bourgeois, N., Rebaudet, S., Piarroux, R., Mauffrey, J.-F., & Ranque, S. (2016). Opportunistic fungal pathogen *Candida glabrata* circulates between humans and yellow-legged gulls. *Scientific Reports, 6,* 36157. 10.1038/srep36157

Albalat, R., & Cañestro, C. (2016). Evolution by gene loss. *Nature Reviews Genetics, 17(7),* 379–391. 10.1038/nrg.2016.39

Alby, K., & Bennett, R. J. (2010). Sexual reproduction in the *Candida* clade: cryptic cycles, diverse mechanisms, and alternative functions. *Cellular and Molecular Life Sciences, 67(19),* 3275–3285. 10.1007/s00018-010-0421-8

Aminnejad, M., Diaz, M., Arabatzis, M., Castañeda, E., Lazera, M., Velegraki, A., Marriott, D., Sorrell, T. C., & Meyer, W. (2012). Identification of novel hybrids between *Cryptococcus neoformans* var. grubii VNI and *Cryptococcus gattii* VGII. *Mycopathologia, 173(5-6),* 337–346. 10.1007/s11046-011-9491-x

Anderson, M. Z., Thomson, G. J., Hirakawa, M. P., & Bennett, R. J. (2019). A "parameiosis" drives depolyploidization and homologous recombination in *Candida albicans. Nature Communications, 10(1),* 4388. 10.1038/s41467-019-12376-2

Arendrup, M. C., & Patterson, T. F. (2017). Multidrug-resistant *Candida*: epidemiology, molecular mechanisms, and treatment. *The Journal of Infectious Diseases, 216(3),* S445–S451. 10.1093/infdis/jix131

Arendrup, M. C., Meletiadis, J., Mouton, J. W., Lagrou, K., Hamal, P., Guinea, J., & the Subcommittee on Antifungal Susceptibility Testing (AFST) of the ESCMID European Committee for Antimicrobial Susceptibility Testing (EUCAST). (2017). Method for the Determination of Broth Dilution Minimum Inhibitory Concentrations of Antifungal Agents For Yeats. EUCAST E.DEF 7.3.1. Available at: http://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/AFST/Files/EUCAST_E_Def_7_3_1_Yeast_testing__definitive.pdf (accessed January 2017).

Azor, M., Gene, J., Cano, J., & Guarro, J. (2007). Universal *In Vitro* Antifungal Resistance of Genetic Clades of the *Fusarium solani* Species Complex. *Antimicrobial Agents and Chemotherapy, 51(4),* 1500-1503. 10.1128/aac.01618-06

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: a new genome

assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *19(5)*, 455–477. 10.1089/cmb.2012.0021

Barnett, J. A. (2008). A history of research on yeasts 12: medical yeasts part 1, *Candida albicans*. *Yeast*, *25(6)*, 385–417. 10.1002/yea.1595

Barreto, F. S., & Burton, R. S. (2013). Evidence for Compensatory Evolution of Ribosomal Proteins in Response to Rapid Divergence of Mitochondrial rRNA. *Molecular Biology and Evolution*, *30(2)*, 310–314. 10.1093/molbev/mss228

Bastos, R. W., Valero, C., Silva, L. P., Schoen, T., Drott, M., Brauer, V., Silva-Rocha, R., Lind, A., Steenwyk, J. L., Rokas, A., Rodrigues, F., Resendiz-Sharpe, A., Lagrou, K., Marcet-Houben, M., Gabaldón, T., McDonnell, E., Reid, I., Tsang, A., Oakley, B. R., Loures, F. V., Almeida, F., Huttenlocher, A., Keller, N. P., Ries, L. N. A., & Goldman, G. H. (2020). Functional Characterization of Clinical Isolates of the Opportunistic Fungal Pathogen *Aspergillus nidulans*. *mSphere*, *5(2)*, e00153. 10.1128/mSphere.00153-20

Bateson, W. (1909). Heredity and Variation in Modern Lights. In A. C. Seward (Ed.), Darwin and Modern Science (pp. 85–101). Cambridge University Press.

Belloch, C., Pérez-Torrado, R., González, S. S., Pérez-Ortín, J. E., García-Martínez, J., Querol, A., & Barrio, E. (2009). Chimeric genomes of natural hybrids of *Saccharomyces cerevisiae* and *Saccharomyces kudriavzevii*. *Applied and Environmental Microbiology*, *75(8)*, 2534–2544. 10.1128/AEM.02282-08

Bennett, R. J., & Johnson, A. D. (2003). Completion of a parasexual cycle in *Candida albicans* by induced chromosome loss in tetraploid strains. *The EMBO Journal*, *22(10)*, 2505–2515. 10.1093/emboj/cdg235

Bennett, R. J. (2015). The parasexual lifestyle of Candida *albicans*. *Current Opinion in Microbiology*, *28*, 10–17. 10.1016/j.mib.2015.06.017

Bensasson, D., Dicks, J., Ludwig, J. M., Bond, C. J., Elliston, A., Roberts, I. N., & James, S. A. (2019). Diverse Lineages of Live *Candida albicans* on Old Oaks. *Genetics, 211(1),* 277–288. 10.1534/genetics.118.301482

Berman, J., & Hadany, L. (2012). Does stress induce (para)sex? Implications for *Candida albicans* evolution. *Trends in Genetics, 28(5),* 197–203. 10.1016/j.tig.2012.01.004

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Putz, J., Middendorf, M., & Stadler, P. F. (2013). MITOS: improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution, 69,* 313–319. 10.1016/j.ympev.2012. 08.023

Bevan, M. W., Uauy, C., Wulff, B. B. H., Zhou, J., Krasileva, K., & Clark, M. D. (2017). Genomic innovation for crop improvement. *Nature, 543(7645),* 346–354. 10.1038/nature22011

Bitar, D., Lortholary, O., Le Strat, Y., Nicolau, J., Coignard, B., Tattevin, P., Che, D., & Dromer, F. (2014). Population-Based Analysis of Invasive Fungal Infections, France, 2001–2010. *Emerging Infectious Diseases, 20(7*), 1163–1169. 10.3201/eid2007.140087

Blanco-Blanco, M. T., Gómez-García, A. C., Hurtado, C., Galán-Ladero, M. A., Lozano, M. del C., García-Tapias, A., & Blanco, M. T. (2014). *Candida orthopsilosis* fungemias in a Spanish tertiary care hospital: incidence, epidemiology and antifungal susceptibility. *Revista Iberoamericana de Micologia, 31(2),* 145–148.

Boekhout, T., Theelen, B., Diaz, M., Fell, J. W., Hop, W. C. J., Abeln, E. C. A., Dromer, F., & Meyer, W. (2001). Hybrid genotypes in the pathogenic yeast *Cryptococcus neoformans*. *Microbiology, 147(4),* 891–907. 10.1099/00221287-147-4-891

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics, 30(15),* 2114–2120. 10.1093/bioinformatics/btu170

Borman, A. M., Muller, J., Walsh-Quantick, J., Szekely, A., Patterson, Z., Palmer, M. D., Fraser, M., & Johnson, E. M. (2019). Fluconazole Resistance in Isolates of Uncommon Pathogenic Yeast Species from the United Kingdom. *Antimicrobial Agents and Chemotherapy, 63(8),* e00211-19. 10.1128/AAC.00211-19

Bovers, M., Hagen, F., Kuramae, E. E., Hoogveld, H. L., Dromer, F., St-Germain, G., & Boekhout, T. (2008). AIDS patient death caused by novel *Cryptococcus neoformans* x *C. gattii* hybrid. *Emerging infectious diseases, 14(7*), 1105–1108. 10.3201/eid1407.080122

Bratton, E. W., El Husseini, N., Chastain, C. A., Lee, M. S., Poole, C., Stürmer, T., Juliano, J. J., Weber, D. J., & Perfect, J. R. (2012). Comparison and temporal trends of three groups with cryptococcosis: HIV-infected, solid organ transplant, and HIV-negative/non-transplant. *PloS One, 7(8),* e43582. 10.1371/journal.pone.0043582

Bretagne, S., Renaudat, C., Desnos-Ollivier, M., Sitbon, K., Lortholary, O., Dromer, F., & French Mycosis Study Group. (2017). Predisposing factors and outcome of uncommon yeast species-related fungaemia based on an exhaustive surveillance programme (2002-14). *The Journal of Antimicrobial Chemotherapy, 72,* 1784–1793. 10.1093/jac/dkx045

Brion, C., Legrand, S., Peter, J., Caradec, C., Pflieger, D., Hou, J., Friedrich, A., Llorente, B., & Schacherer, J. (2017). Variation of the meiotic recombination landscape and properties over a broad evolutionary distance in yeasts. *PLoS Genetics, 13(8),* e1006917. 10.1371/journal.pgen.1006917

Brisse, S., Pannier, C., Angoulvant, A., de Meeus, T., Diancourt, L., Faure, O., Muller, H., Peman, J., Viviani, M. A., Grillot, R., Dujon, B., Fairhead, C., & Hennequin, C. (2009). Uneven distribution of mating types among genotypes of *Candida glabrata* isolates from clinical samples. *Eukaryotic Cell, 8(3),* 287–295. 10.1128/EC.00215-08

Brown, G. D., Denning, D. W., Gow, N. A. R., Levitz, S. M., Netea, M. G., & White, T. C. (2012). Hidden Killers: Human Fungal

Infections. *Science Translational Medicine, 4(165),* 165rv13. 10.1126/scitranslmed.3004404

Brunke, S., & Hube, B. (2013). Two unlike cousins: *Candida albicans* and *C. glabrata* infection strategies. *Cellular Microbiology, 15(5),* 701–708. 10.1111/cmi.12091

Butler, G., Kenny, C., Fagan, A., Kurischko, C., Gaillardin, C., & Wolfe, K. H. (2004). Evolution of the *MAT* locus and its Ho endonuclease in yeast species. *Proceedings of the National Academy of Sciences of the United States of America, 101(6),* 1632–1637. 10.1073/pnas.0304170101

Butler, G., Rasmussen, M. D., Lin, M. F., Santos, M. A. S., Sakthikumar, S., Munro, C. A., Rheinbay, E., Grabherr, M., Forche, A., Reedy, J. L., Agrafioti, I., Arnaud, M. B., Bates, S., Brown, A. J. P., Brunke, S., Costanzo, M. C., Fitzpatrick, D. A., de Groot, P. W. J., Harris, D., … Cuomo, C. A. (2009). Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature, 459(7247),* 657–662. 10.1038/nature08064

Byrnes, E. J., & Heitman, J. (2009). *Cryptococcus gattii* outbreak expands into the Northwestern United States with fatal consequences. *F1000 Biology Reports, 1,* 62. 10.3410/B1-62

Byrnes, E. J., & Marr, K. A. (2011). The Outbreak of *Cryptococcus gattii* in Western North America: Epidemiology and Clinical Issues. *Current Infectious Disease Reports, 13(3),* 256–261. 10.1007/s11908-011-0181-0

Bzymek, M., Thayer, N. H., Oh, S. D., Kleckner, N., & Hunter, N. (2010). Double Holliday junctions are intermediates of DNA break repair. *Nature, 464(7290),* 937–941. 10.1038/nature08868

Callon, C., Duthoit, F., Delbès, C., Ferrand, M., Le Frileux, Y., De Crémoux, R., & Montel, M.-C. (2007). Stability of microbial communities in goat milk during a lactation year: molecular approaches. *Systematic and Applied Microbiology, 30,* 547–560. 10.1016/j. syapm.2007.05.004

Cantón, E., Pemán, J., Quindós, G., Eraso, E., Miranda-Zapico, I., Álvarez, M., Merino, P., Campos-Herrero, I., Marco, F., de la Pedrosa, E. G. G., Yagüe, G., Guna, R., Rubio, C., Miranda, C., Pazos, C., Velasco, D., & the FUNGEMYCA Study Group. (2011). Prospective Multicenter Study of the Epidemiology, Molecular Identification, and Antifungal Susceptibility of *Candida parapsilosis*, *Candida orthopsilosis*, and *Candida metapsilosis* Isolated from Patients with Candidemia. *Antimicrobial Agents and Chemotherapy*, *55(12)*, 5590–5596. 10.1128/aac.00466-11

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25(15)*, 1972–1973. 10.1093/bioinformatics/btp348

Capella-Gutierrez, S., Kauff, F., & Gabaldón, T. (2014). A phylogenomics approach for selecting robust sets of phylogenetic markers. *Nucleic Acids Research*, *42(7)*, e54. 10.1093/nar/gku071

Carneiro, M. O., Russ, C., Ross, M. G., Gabriel, S. B., Nusbaum, C., & DePristo, M. A. (2012). Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC Genomics*, *13*, 375. 10.1186/1471-2164-13-375

Carreté, L., Ksiezopolska, E., Pegueroles, C., Gómez-Molero, E., Saus, E., Iraola-Guzmán, S., Loska, D., Bader, O., Fairhead, C., & Gabaldón, T. (2018). Patterns of Genomic Variation in the Opportunistic Pathogen *Candida glabrata* Suggest the Existence of Mating and a Secondary Association with Humans. *Current Biology*, *28(1)*, 15–27.e7. 10.1016/j.cub.2017.11.027

Casadevall, A. (2012). Amoeba provide insight into the origin of virulence in pathogenic fungi. *Advances in Experimental Medicine and Biology*, *710*, 1–10. 10.1007/978-1-4419-5638-5_1

Cauchie, M., Desmet, S., & Lagrou, K. (2017). *Candida* and its dual lifestyle as a commensal and a pathogen. *Research in Microbiology*, *168(9-10)*, 802–810. 10.1016/j.resmic.2017.02.005

Cendejas-Bueno, E., Gomez-Lopez, A., Mellado, E., Rodriguez-Tudela, J. L., & Cuenca-Estrella, M. (2010). Identification of

pathogenic rare yeast species in clinical samples: comparison between phenotypical and molecular methods. *Journal of Clinical Microbiology, 48,* 1895–1899. 10.1128/JCM.00336-10

Chang, C.-F., Lin, Y.-C., Chen, S.-F., Carvajal Barriga, E. J., Barahona, P. P., James, S. A., Bond, C. J., Roberts, I. N., & Lee, C.-F. (2012). *Candida theae* sp. nov., a new anamorphic beverage-associated member of the Lodderomyces clade. *International Journal of Food Microbiology, 153(1-2),* 10–14. 10.1016/j.ijfoodmicro.2011.09.012

Chen, Z. J. (2013). Genomic and epigenetic insights into the molecular bases of heterosis. *Nature Reviews Genetics, 14(7),* 471–482. 10.1038/nrg3503

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Conception, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C., Ecker, J. R., Cantu, D., Rank, D. R., & Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods, 13(12),* 1050–1054. 10.1038/nmeth.4035

Chow, B. D. W., Linden, J. R., & Bliss, J. M. (2012). *Candida parapsilosis* and the neonate: epidemiology, virulence and host defense in a unique patient setting. *Expert Review of Anti-infective Therapy, 10(8),* 935–946. 10.1586/eri.12.74

Cogliati, M., D'Amicis, R., Zani, A., Montagna, M. T., Caggiano, G., De Giglio, O., Balbino, S., De Donno, A., Serio, F., Susever, S., Ergin, C., Velegraki, A., Ellabib, M. S., Nardoni, S., Macci, C., Oliveri, S., Trovato, L., Dipineto, L., Rickerts, V., … Colom, M. F. (2016). Environmental distribution of *Cryptococcus neoformans* and *C. gattii* around the Mediterranean basin. *FEMS Yeast Research, 16(4),* fow045. 10.1093/femsyr/fow045

Comai L. (2005). The advantages and disadvantages of being polyploid. *Nature reviews genetics, 6(11),* 836–846. 10.1038/nrg1711

Consortium OPATHY, & Gabaldón, T. (2019). Recent trends in molecular diagnostics of yeast infections: from PCR to NGS. *FEMS Microbiology Reviews, 43(5)*, 517–547. 10.1093/femsre/fuz015

Crawshaw, W. M., MacDonald, N. R., & Duncan, G. (2005). Outbreak of *Candida rugosa* mastitis in a dairy herd after intramammary antibiotic treatment. *The Veterinary Record, 156*, 812–813. 10.1136/vr.156.25.812

Cuomo, C. A., Shea, T., Yang, B., Rao, R., & Forche, A. (2017). Whole genome sequence of the heterozygous clinical isolate *Candida krusei*. *G3: Genes|Genomes|Genetics, 7(9)*, 2883–2889. 10.1534/g3.117.043547

D'Antonio, D., Violante, B., Mazzoni, A., Bonfini, T., Capuani, M. A., D'Aloia, F., Iacone, A., Schioppa, F., & Romano, F. (1998). A nosocomial cluster of *Candida inconspicua* infections in patients with hematological malignancies. *Journal of Clinical Microbiology, 36(3)*, 792–795.

D'Souza, C. A., Kronstad, J. W., Taylor, G., Warren, R., Yuen, M., Hu, G., Jung, W. H., Sham, A., Kidd, S. E., Tangen, K., Lee, N., Zeilmaker, T., Sawkins, J., McVicker, G., Shah, S., Gnerre, S., Griggs, A., Zeng, Q., Bartlett, K., … Cuomo, C. A. (2011). Genome Variation in *Cryptococcus gattii*, an Emerging Pathogen of Immunocompetent Hosts. *mBio, 2(1)*, e00342. 10.1128/mbio.00342-10

Dagilis, A. J., Kirkpatrick, M., & Bolnick, D. I. (2019). The evolution of hybrid fitness during speciation. *PLoS Genetics, 15(5)*, e1008125. 10.1371/journal.pgen.1008125

Dayani, Y., Simchen, G., & Lichten, M. (2011). Meiotic recombination intermediates are resolved with minimal crossover formation during return-to-growth, an analogue of the mitotic cell cycle. *PLoS Genetics, 7(5)*, e1002083. 10.1371/journal.pgen.1002083

Deonovic, B., Wang, Y., Weirather, J., Wanf, X.-J-, & Au, K. F. (2017) IDP-ASE: haplotyping and quantifying allele-specific

expression at the gene and gene isoform level by hybrid sequencing. *Nucleic Acids Research, 45,* e32. 10.1093/nar/gkw1076

Depotter, J. R. L., Seidl, M. F., Wood, T. A., & Thomma, B. P. (2016). Interspecific hybridization impacts host range and pathogenicity of filamentous microbes. *Current Opinion in Microbiology, 32,* 7–13. 10.1016/j.mib.2016.04.005

Diekema, D., Arbefeville, S., Boyken, L., Kroeger, J., & Pfaller, M. (2012). The changing epidemiology of healthcare-associated candidemia over three decades. *Diagnostic Microbiology and Infectious Disease, 73(1),* 45–48. 10.1016/j.diagmicrobio.2012.02.001

Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research, 45,* e18. 10.1093/nar/gkw955

Diezmann, S., & Dietrich, F. S. (2009). *Saccharomyces cerevisiae*: population divergence and resistance to oxidative stress in clinical, domesticated and wild isolates. *PloS one, 4(4),* e5317. 10.1371/journal.pone.0005317

Dobzhansky, T. (1934). Studies on hybrid sterility: I. Spermatogenesis in pure and hybrid Drosophila pseudoobscura. *Zeitschrift Fur Zellforschung Und Mikroskopische Anatomie, 21(2),* 169–223.

Douglass, A. P., Offei, B., Braun-Galleani, S., Coughlan, A. Y., Martos, A. A. R., Ortiz-Merino, R. A., Byrne, K. P., & Wolfe, K. H. (2018). Population genomics shows no distinction between pathogenic *Candida krusei* and environmental *Pichia kudriavzevii*: one species, four names. *PLoS Pathogens, 14,* e1007138. 10.1371/journal.ppat.1007138

Edge, P., Bafna, V., & Bansal, V. (2017). HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Research, 27(5),* 801–812. 10.1101/gr.213462.116

Egue, L. A. N., Bouatenin, J.-P. K. M., N'guessan, F. K., & Koussemon-Camara, M. (2018). Virulence factors and

determination of antifungal susceptibilities of *Candida* species isolated from palm wine and sorghum beer. *Microbial Pathogenesis, 124,* 5–10. 10.1016/j.micpath.2018.08.007

Ene, I. V., Farrer, R. A., Hirakawa, M. P., Agwamba, K., Cuomo, C. A., & Bennett, R. J. (2018). Global analysis of mutations driving microevolution of a heterozygous diploid fungal pathogen. *Proceedings of the National Academy of Sciences of the United States of America, 115(37),* E8688–E8697. 10.1073/pnas.1806002115

Fabre, E., Muller, H., Therizols, P., Lafontaine, I., Dujon, B., & Fairhead, C. (2005). Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing, and subtelomeres. *Molecular Biology and Evolution, 22(4),* 856–873. 10.1093/molbev/msi070

Finkel, J. S., & Mitchell, A. P. (2011). Genetic control of *Candida albicans* biofilm development. *Nature Reviews Microbiology, 9(2),* 109–118. 10.1038/nrmicro2475

Fonseca, D. M., Keyghobadi, N., Malcolm, C. A., Mehmet, C., Schaffner, F., Mogi, M., Fleischer, R. C., & Wilkerson, R. C. (2004). Emerging vectors in the *Culex pipiens* complex. *Science, 303(5663),* 1535–1538. 10.1126/science.1094247

Forche, A., Alby, K., Schaefer, D., Johnson, A. D., Berman, J., & Bennett, R. J. (2008). The parasexual cycle in *Candida albicans* provides an alternative pathway to meiosis for the formation of recombinant strains. *PLoS Biology, 6(5),* e110. 10.1371/journal.pbio.0060110

Forsberg, K., Woodworth, K., Walters, M., Berkow, E. L., Jackson, B., Chiller, T., & Vallabhaneni, S. (2019). *Candida auris*: the recent emergence of a multidrug-resistant fungal pathogen. *Medical Mycology, 57,* 1–12. 10.1093/mmy/myy054

Fricova, D., Valach, M., Farkas, Z., Pfeiffer, I., Kucsera, J., Tomaska, L., & Nosek, J. (2010). The mitochondrial genome of the pathogenic yeast *Candida subhashii*: GC-rich linear DNA with a protein covalently attached to the 5’ termini. *Microbiology, 156(7),* 2153–2163. 10.1099/mic.0.038646-0

Frisch, M., Thiemann, A., Fu, J., Schrag, T. A., Scholten, S., & Melchinger, A. E. (2010). Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize. *Theoretical and Applied Genetics, 120(2),* 441–450. 10.1007/s00122-009-1204-1

Gabaldón, T. (2008). Large-scale assignment of orthology: back to phylogenetics? *Genome Biology, 9,* 235. 10.1186/gb-2008-9-10-235

Gabaldón, T., Martin, T., Marcet-Houben, M., Durrens, P., Bolotin-Fukuhara, M., Lespinet, O., Arnaise, S., Boisnard, S., Aguileta, G., Atanasova, R., Bouchier, C., Couloux, A., Creno, S., Almeida Cruz, J., Devillers, H., Enache-Angoulvant, A., Guitard, J., Jaouen, L., Ma, L., … Fairhead, C. (2013). Comparative genomics of emerging pathogens in the *Candida glabrata* clade. *BMC Genomics, 14,* 623. 10.1186/1471-2164-14-623

Gabaldón, T., & Carreté, L. (2016). The birth of a deadly yeast: tracing the evolutionary emergence of virulence traits in *Candida glabrata*. *FEMS Yeast Research, 16(2),* fov110. 10.1093/femsyr/fov110

Gabaldón, T., Naranjo-Ortíz, M. A., & Marcet-Houben, M. (2016). Evolutionary genomics of yeast pathogens in the Saccharomycotina. *FEMS Yeast Research, 16(6),* fow064. 10.1093/femsyr/fow064

Gabaldón, T., & Fairhead, C. (2019). Genomes shed light on the secret life of *Candida glabrata*: not so asexual, not so commensal. *Current Genetics, 65(1),* 93–98. 10.1007/s00294-018-0867-z

Gácser, A., Schäfer, W., Nosanchuk, J. S., Salomon, S., & Nosanchuk, J. D. (2007). Virulence of *Candida parapsilosis*, *Candida orthopsilosis*, and *Candida metapsilosis* in reconstituted human tissue models. *Fungal Genetics and Biology, 44(12),* 1336–1341. 10.1016/j.fgb.2007.02.002

Galanis, E., MacDougall, L., Kidd, S., Morshed, M., & the British Columbia Cryptococcus gattii Working Group. (2010). Epidemiology of *Cryptococcus gattii*, British Columbia, Canada,

1999–2007. *Emerging Infectious Diseases, 16(2),* 251–257. 10.3201/eid1602.090900

Galocha, M., Pais, P., Cavalheiro, M., Pereira, D., Viana, R., & Teixeira, M. C. (2019). Divergent Approaches to Virulence in *C. albicans* and *C. glabrata:* Two Sides of the Same Coin. *International Journal of Molecular Sciences, 20(9),* 2345. 10.3390/ijms20092345

Garcia-Effron, G., Katiyar, S. K., Park, S., Edlind, T. D., & Perlin, D. S. (2008). A Naturally Occurring Proline-to-Alanine Amino Acid Change in *Fks1p* in *Candida parapsilosis*, *Candida orthopsilosis*, and *Candida metapsilosis* Accounts for Reduced Echinocandin Susceptibility. *Antimicrobial Agents and Chemotherapy*, *52(7),* 2305–2312. 10.1128/aac.00262-08

Garrison, E. & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv*, arXiv: 1207.3907.

Gibson, B., & Liti, G. (2015). *Saccharomyces pastorianus*: genomic insights inspiring innovation for industry. *Yeast, 32(1),* 17–27. 10.1002/yea.3033

Gladieux, P., Ropars, J., Badouin, H., Branca, A., Aguileta, G., de Vienne, D. M., Rodríguez de la Vega, R. C., Branco, S., & Giraud, T. (2014). Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. *Molecular Ecology*, *23(4),* 753–773. 10.1111/mec.12631

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics, 17(6),* 333–351. 10.1038/nrg.2016.49

Goyer, M., Loiselet, A., Bon, F., L'Ollivier, C., Laue, M., Holland, G., Bonnin, A., & Dalle, F. (2016). Intestinal Cell Tight Junctions Limit Invasion of *Candida albicans* through Active Penetration and Endocytosis in the Early Stages of the Interaction of the Fungus with the Intestinal Barrier. *PloS One, 11(3),* e0149159. 10.1371/journal.pone.0149159

Guitard, J., Angoulvant, A., Letscher-Bru, V., L'Ollivier, C., Cornet, M., Dalle, F., Grenouillet, F., Lacroix, C., Vekhoff, A., Maury, E., Caillot, D., Charles, P. E., Pili-Floury, S., Herbrecht, R., Raffoux, E., Brethon, B., & Hennequin, C. (2013). Invasive infections due to *Candida norvegensis* and *Candida inconspicua*: report of 12 cases and review of the literature. Medical Mycology. Official Publication of the International Society for Human and Animal Mycology. *Medical Mycology, 51(8),* 795–799. 10.3109/13693786.2013.807444

Guitard, J., Atanasova, R., Brossas, J. Y., Meyer, I., Gits, M., Marinach, C., Vellaissamy, S., Angoulvant, A., Mazier, D., & Hennequin, C. (2015). *Candida inconspicua* and *Candida norvegensis*: new insights into identification in relation to sexual reproduction and genome organization. *Journal of Clinical Microbiology, 53(5),* 1655–1661. 10.1128/JCM.02913-14

Gunsilius, E., Lass-Flörl, C., Kähler, C. M., Gastl, G., & Petzer, A. L. (2001). *Candida ciferrii*, a new fluconazole-resistant yeast causing systemic mycosis in immunocompromised patients. *Annals of Hematology, 80(3),* 178–179. 10.1007/s002770000252

Guo, M., Rupe, M. A., Yang, X., Crasta, O., Zinselmeier, C., Smith, O. S., & Bowen, B. (2006). Genome-wide transcript analysis of maize hybrids: allelic additive gene expression and yield heterosis. *Theoretical and Applied Genetics, 13(5),* 831–845. 10.1007/s00122-006-0335-x

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics, 29(8),* 1072–1075. 10.1093/bioinformatics/btt086

Haber, J. E. (2012). Mating-type genes and *MAT* switching in *Saccharomyces cerevisiae*. *Genetics, 191(1),* 33–64. 10.1534/genetics.111.134577

Hagen, F., Khayhan, K., Theelen, B., Kolecka, A., Polacheck, I., Sionov, E., Falk, R., Parnmen, S., Lumbsch, H. T., & Boekhout, T. (2015). Recognition of seven species in the *Cryptococcus gattii/Cryptococcus neoformans* species complex. *Fungal Genetics and Biology,* 78, 16–48. 10.1016/j.fgb.2015.02.009

He, G., Zhu, X., Elling, A. A., Chen, L., Wang, X., Guo, L., Liang, M., He, H., Zhang, H., Chen, F., Qi, Y., Chen, R., & Deng, X.-W. (2010). Global Epigenetic and Transcriptional Trends among Two Rice Subspecies and Their Reciprocal Hybrids. *The Plant Cell*, *22(1)*, 17–33. 10.1105/tpc.109.072041

Henao-Martínez, A. F., & Beckham, J. D. (2015). Cryptococcosis in solid organ transplant recipients. *Current Opinion in Infectious Diseases*, *28(4)*, 300–307. 10.1097/qco.0000000000000171

Heymann, P., Gerads, M., Schaller, M., Dromer, F., Winkelmann, G., & Ernst, J. F. (2002). The siderophore iron transporter of *Candida albicans* (*Sit1p/Arn1p*) mediates uptake of ferrichrome-type siderophores and is required for epithelial invasion. *Infection and Immunity*, *70(9)*, 5246–5255. 10.1128/IAI.70.9.5246-5255.2002

Hickman, M. A., Zeng, G., Forche, A., Hirakawa, M. P., Abbey, D., Harrison, B. D., Wang, Y.-M., Su, C.-H., Bennett, R. J., Wang, Y., & Berman, J. (2013). The "obligate diploid" *Candida albicans* forms mating-competent haploids. *Nature*, *494(7435)*, 55–59. 10.1038/nature11865

Hilber-Bodmer, M., Schmid, M., Ahrens, C. H., & Freimoser, F. M. (2017). Competition assays and physiological experiments of soil and phyllosphere yeasts identify *Candida subhashii* as a novel antagonist of filamentous fungi. *BMC Microbiology*, *17(1)*, 4. 10.1186/s12866-016-0908-z

Hirakawa, M. P., Martinez, D. A., Sakthikumar, S., Anderson, M. Z., Berlin, A., Gujja, S., Zeng, Q., Zisson, E., Wang, J. M., Greenberg, J. M., Berman, J., Bennett, R. J., & Cuomo, C. A. (2015). Genetic and phenotypic intra-species variation in *Candida albicans*. *Genome Research*, *25(3)*, 413–425. 10.1101/gr.174623.114

Ho, J., Wickramasinghe, D. N., Nikou, S.-A., Hube, B., Richardson, J. P., & Naglik, J. R. (2020). Candidalysin Is a Potent Trigger of Alarmin and Antimicrobial Peptide Release in Epithelial Cells. *Cells*, *9(3)*, 699. 10.3390/cells9030699

Holzheimer, R. G., and Dralle, H. (2002). Management of mycoses in surgical patients - review of the literature. *European Journal of Medical Research, 7,* 200–226.

Hou, J., & Schacherer, J. (2016). Negative epistasis: a route to intraspecific reproductive isolation in yeast? *Current Genetics, 62(1),* 25–29. 10.1007/s00294-015-0505-y

Hovhannisyan, H., Saus, E., Ksiezopolska, E., & Gabaldón, T. (2020). The transcriptional aftermath in two independently formed hybrids of the opportunistic pathogen *Candida orthopsilosis*. *mSphere, 5(3),* e00282. 10.1128/mSphere.00282-20

Hoyer, L. L., Green, C. B., Oh, S.-H., & Zhao, X. (2008). Discovering the secrets of the *Candida albicans* agglutinin-like sequence (ALS) gene family--a sticky pursuit. *Medical Mycology, 46(1),* 1–15. 10.1080/13693780701435317

Hu, G., Cheng, P.-Y., Sham, A., Perfect, J. R., & Kronstad, J. W. (2008). Metabolic adaptation in *Cryptococcus neoformans* during early murine pulmonary infection. *Molecular Microbiology, 69(6), 1456*–1475. 10.1111/j.1365-2958.2008.06374.x

Huang, M., & Hull, C. M. (2017). Sporulation: how to survive on planet Earth (and beyond). *Current Genetics, 63(5),* 831–838. 10.1007/s00294-017-0694-7

Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L. P., Marcet-Houben, M., & Gabaldón, T. (2014). PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Research, 42,* D897–D902. 10.1093/nar/ gkt1177

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution, 34(8),* 2115–2122. 10.1093/molbev/msx148

Hull, C. M., & Johnson, A. D. (1999). Identification of a mating type-like locus in the asexual pathogenic yeast *Candida albicans*. *Science, 285(5431),* 1271–1275. 10.1126/science.285.5431.1271

Hull, C. M., Raisner, R. M., & Johnson, A. D. (2000). Evidence for mating of the "asexual" yeast *Candida albicans* in a mammalian host. *Science, 289(5477),* 307–310. 10.1126/science.289.5477.307

Hunter, N., Chambers, S. R., Louis, E. J., & Borts, R. H. (1996). The mismatch repair system contributes to meiotic sterility in an interspecific yeast hybrid. *The EMBO Journal, 15(7),* 1726–1733.

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology Evolution, 23,* 254–267. 10.1093/molbev/msj030

Hwang, S., Kim, E., Lee, I., & Marcotte, E. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports, 5(1),* 47. 10.1038/srep17875

Inderbitzin, P., Bostock, R. M., Davis, R. M., Usami, T., Platt, H. W., & Subbarao, K. V. (2011a). Phylogenetics and taxonomy of the fungal vascular wilt pathogen *Verticillium*, with the descriptions of five new species. *PloS One, 6(12),* e28341. 10.1371/journal.pone.0028341

Inderbitzin, P., Davis, R. M., Bostock, R. M., & Subbarao, K. V. (2011b). The ascomycete *Verticillium longisporum* is a hybrid and a plant pathogen with an expanded host range. *PloS One, 6(3),* e18260. 10.1371/journal.pone.0018260

Iraqui, I., Garcia-Sanchez, S., Aubert, S., Dromer, F., Ghigo, J.-M., D'Enfert, C., & Janbon, G. (2004). The *Yak1p* kinase controls expression of adhesins and biofilm formation in *Candida glabrata* in a Sir4p-dependent pathway. *Molecular Microbiology, 55(4),* 1259–1271. 10.1111/j.1365-2958.2004.04475.x

Jackson, A. P., Gamble, J. A., Yeomans, T., Moran, G. P., Saunders, D., Harris, D., Aslett, M., Barrell, J. F., Butler, G., Citiulo, F., Coleman, D. C., de Groot, P. W. J., Goodwin, T. J., Quail, M. A., McQuillan, J., Munro, C. A., Pain, A., Poulter, R. T., Rajandream, M.-A., … Berriman, M. (2009). Comparative genomics of the fungal pathogens *Candida dubliniensis* and *Candida albicans*. *Genome Research, 19(12),* 2231–2244. 10.1101/gr.097501.109

Janbon, G., Ormerod, K. L., Paulet, D., Byrnes, E. J., 3rd, Yadav, V., Chatterjee, G., Mullapudi, N., Hon, C.-C., Billmyre, R. B., Brunel, F., Bahn, Y.-S., Chen, W., Chen, Y., Chow, E. W. L., Coppée, J.-Y., Floyd-Averette, A., Gaillardin, C., Gerik, K. J., Goldberg, J., … Dietrich, F. S. (2014). Analysis of the genome and transcriptome of *Cryptococcus neoformans* var. grubii reveals complex RNA expression and microevolution leading to virulence attenuation. *PLoS Genetics*, *10(4)*, e1004261. 10.1371/journal.pgen.1004261

Jeffries, T. W., Grigoriev, I. V., Grimwood, J., Laplaza, J. M., Aerts, A., Salamov, A., Schmutz, J., Lindquist, E., Dehal, P., Shapiro, H., Jin, Y. S., Passoth, V., & Richardson, P. M. (2007). Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nature biotechnology*, *25(3)*, 319–326. 10.1038/nbt1290

Jhuang, H.-Y., Lee, H.-Y., & Leu, J.-Y. (2017). Mitochondrial-nuclear co-evolution leads to hybrid incompatibility through pentatricopeptide repeat proteins. *EMBO Reports*, *18(1)*, 87–101. 10.15252/embr.201643311

Ji, Z.-H., Jia, J. H., & Bai, F.-Y. (2009). Four novel *Candid*a species in the *Candida albicans/Lodderomyces elongisporus* clade isolated from the gut of flower beetles. *Antonie van Leeuwenhoek*, *95(1)*, 23–32. 10.1007/s10482-008-9282-7

Johnson, S. M., Carlson, E. L., & Pappagianis, D. (2015). *Coccidioides* species determination: does sequence analysis agree with restriction fragment length polymorphism?. *Mycopathologia*, *179(5-6)*, 373–379. 10.1007/s11046-014-9857-y

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30(9)*, 1236–1240. 10.1093/bioinformatics/btu031

Jordà-Marcos, R., Alvarez-Lerma, F., Jurado, M., Palomar, M., Nolla-Salas, J., León, M. A., León, C., & EPCAN Study Group. (2007). Risk factors for candidaemia in critically ill patients: a prospective surveillance study. *Mycoses, 50(4),* 302–310. 10.1111/j.1439-0507.2007.01366.x

Jung, D. S., Farmakiotis, D., Jiang, Y., Tarrand, J. J., & Kontoyiannis, D. P. (2015). Uncommon *Candida* Species Fungemia among Cancer Patients, Houston, Texas, USA. *Emerging infectious diseases, 21(11),* 1942–1950. 10.3201/eid2111.150404

Kadyk, L. C., & Hartwell, L. H. (1992). Sister chromatids are preferred over homologs as substrates for recombinational repair in *Saccharomyces cerevisiae*. *Genetics, 132(2),* 387–402.

Kaewkrajay, C., Chanmethakul, T., & Limtong, S. (2020). Assessment of Diversity of Culturable Marine Yeasts Associated with Corals and Zoanthids in the Gulf of Thailand, South China Sea. *Microorganisms, 8(4),* 474. 10.3390/microorganisms8040474

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., Kohara, Y., Fujiyama, A., Hayashi, T., & Itoh, T. (2014). Efficient *de novo* assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Research, 24(8),* 1384–1395. 10.1101/gr.170720.113

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution, 30(4),* 772–780. 10.1093/molbev/mst010

Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics, 20(4),* 1160–1166. 10.1093/bib/bbx108

Khunnamwong, P., Lertwattanasakul, N., Jindamorakot, S., Limtong, S., & Lachance, M. A. (2015). Description of *Diutina* gen. nov., *Diutina siamensis*, f.a. sp. nov., and reassignment of *Candida catenulata*, *Candida mesorugosa*, *Candida neorugosa*, *Candida*

pseudorugosa, *Candida ranongensis*, *Candida rugosa* and *Candida scorzettiae* to the genus *Diutina*. *International journal of systematic and evolutionary microbiology*, *65(12)*, 4701–4709. 10.1099/ijsem.0.000634

Khunnamwong, P., Jindamorakot, S., & Limtong, S. (2018). Endophytic yeast diversity in leaf tissue of rice, corn and sugarcane cultivated in Thailand assessed by a culture-dependent approach. *Fungal Biology*, *122(8)*, 785–799. 10.1016/j.funbio.2018.04.006

King, K. C., Stelkens, R. B., Webster, J. P., Smith, D. F., & Brockhurst, M. A. (2015). Hybridization in Parasites: Consequences for Adaptive Evolution, Pathogenesis, and Public Health in a Changing World. *PLoS Pathogens*, *11(9)*, e1005098. 10.1371/journal.ppat.1005098

Kolmogorov, M., Raney, B., Paten, B., & Pham, S. (2014). Ragout - a reference-assisted assembly tool for bacterial genomes. *Bioinformatics*, *30(12)*, i302–i309. 10.1093/bioinformatics/btu280

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., Richard McCombie, W., Jarvis, E. D., & Phillippy, A. M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, *30(7)*, 693–700. 10.1038/nbt.2280

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation. *Genome Research*, *27(5)*, 722–736. 10.1101/gr.215087.116

Krogerus, K., Magalhães, F., Vidgren, V., & Gibson, B. (2017). Novel brewing yeast hybrids: creation and application. *Applied Microbiology and Biotechnology*, *101(1)*, 65–78. 10.1007/s00253-016-8007-5

Krogerus, K., Preiss, R., & Gibson, B. (2018). A Unique Saccharomyces cerevisiae × Saccharomyces uvarum Hybrid Isolated From Norwegian Farmhouse Beer: Characterization and Reconstruction. *Frontiers in microbiology*, *9*, 2253. 10.3389/fmicb.2018.02253

Krukowski, H., Lisowski, A., Rózański, P., & Skórka, A. (2006). Yeasts and algae isolated from cows with mastitis in the south-eastern part of Poland. *Polish journal of veterinary sciences, 9(3),* 181–184.

Ksiezopolska, E., & Gabaldón, T. (2018). Evolutionary Emergence of Drug Resistance in *Candida* Opportunistic Pathogens. *Genes, 9(9),* 461. 10.3390/genes9090461

Kullberg, B. J., & Arendrup, M. C. (2015). Invasive Candidiasis. *New England Journal of Medicine, 373(15),* 1445–1456. 10.1056/nejmra1315399

Kullberg, B. J., Arendrup, M. C. (2015). Invasive Candidiasis. *The New England journal of medicine, 373,* 1445–1456. 10.1056/NEJMra1315399

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome biology, 5(2),* R12. 10.1186/gb-2004-5-2-r12

Kurtzman, C. P., & C. J. Robnett, (2007). Multigene phylogenetic analysis of the *Trichomonascus, Wickerhamiella* and *Zygoascus* yeast clades, and the proposal of *Sugiyamaella* gen. nov. and 14 new species combinations. *FEMS Yeast Research, 7,* 141–151. 10.1111/j.1567- 1364.2006.00157.x

Kurtzman, C. P., Robnett, C. J., & Basehoar-Powers, E. (2008). Phylogenetic relationships among species of *Pichia, Issatchenkia* and *Williopsis* determined from multigene sequence analysis, and the proposal of *Barnettozyma* gen. nov., *Lindnera* gen. nov. and *Wickerhamomyces* gen. nov. *FEMS Yeast Research, 8,* 939–954. 10.1111/j.1567-1364.2008.00419.x

Kurtzman, C., Fell, J. W., & Boekhout, T. (2011). The Yeasts: A Taxonomic Study. Amsterdam: Elsevier.

LaFave, M. C., & Sekelsky, J. (2009). Mitotic recombination: why? when? how? where? *PLoS Genetics, 5(3),* e1000411. 10.1371/journal.pgen.1000411

Lancaster, S. M., Payen, C., Smukowski Heil, C., & Dunham, M. J. (2019). Fitness benefits of loss of heterozygosity in hybrids. *Genome Research, 29(10),* 1685–1692. 10.1101/gr.245605.118

Langdon, Q. K., Peris, D., Kyle, B., & Hittinger, C. T. (2018). sppIDer: A Species Identification Tool to Investigate Hybrid Genomes with High-Throughput Sequencing. *Molecular Biology and Evolution, 35(11),* 2835–2849. 10.1093/molbev/msy166

Langdon, Q. K., Peris, D., Baker, E. P., Opulente, D. A., Nguyen, H.-V., Bond, U., Gonçalves, P., Sampaio, J. P., Libkind, D., & Hittinger, C. T. (2019). Fermentation innovation through complex hybridization of wild and domesticated yeasts. *Nature Ecology & Evolution, 3(11),* 1576–1586. 10.1038/s41559-019-0998-8

Lasheras, A., Rogues, A. M., Peyrere, S., Boulard, G., Bebear, C. M., Gachie, J. P., Bretagne, S., & Dromer, F. (2007). *Candida albicans* outbreak in a neurosurgical intensive care unit. *The Journal of Hospital Infection, 65(2),* 181–182. 10.1016/j.jhin.2006.10.009

Lass-Flörl, C. (2009). The changing face of epidemiology of invasive fungal disease in Europe. *Mycoses, 52(3),* 197–205. 10.1111/j.1439-0507.2009.01691.x

Lattif, A. A., Mukherjee, P. K., Chandra, J., Swindell, K., Lockhart, S. R., Diekema, D. J., Pfaller, M. A., & Ghannoum, M. A. (2010). Characterization of biofilms formed by *Candida parapsilosis, C. metapsilosis,* and *C. orthopsilosis*. *International Journal of Medical Microbiology, 300(4),* 265–270. 10.1016/j.ijmm.2009.09.001

Laureau, R., Loeillet, S., Salinas, F., Bergström, A., Legoix-Né, P., Liti, G., & Nicolas, A. (2016). Extensive Recombination of a Yeast Diploid Hybrid through Meiotic Reversion. *PLoS Genetics, 12(2),* e1005781. 10.1371/journal.pgen.1005781

Leducq, J.-B., Nielly-Thibault, L., Charron, G., Eberlein, C., Verta, J.-P., Samani, P., Sylvester, K., Hittinger, C. T., Bell, G., & Landry, C. R. (2016). Speciation driven by hybridization and chromosomal plasticity in a wild yeast. *Nature Microbiology, 1,* 15003. 10.1038/nmicrobiol.2015.3

Lee, H.-Y., Chou, J.-Y., Cheong, L., Chang, N.-H., Yang, S.-Y., & Leu, J.-Y. (2008). Incompatibility of Nuclear and Mitochondrial Genomes Causes Hybrid Sterility between Two Yeast Species. *Cell, 135(6),* 1065–1073. 10.1016/j.cell.2008.10.047

Lee, Y., Marsden, C. D., Norris, L. C., Collier, T. C., Main, B. J., Fofana, A., Cornel, A. J., & Lanzaro, G. C. (2013). Spatiotemporal dynamics of gene flow and hybrid fitness between the M and S forms of the malaria mosquito, *Anopheles gambiae*. *Proceedings of the National Academy of Sciences of the United States of America, 110(49),* 19854–19859. 10.1073/pnas.1316851110

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics , 25(16),* 2078–2079. 10.1093/bioinformatics/btp352

Li, H. (2011) A statistical framework for SNP calling, mutation discovery, as- sociation mapping and population genetical parameter estimation from sequencing data. *Bioinformatics, 27,* 2987–2993. 10.1093/bioinformatics/btr509

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 1303.3997:1–3.

Li, W., Averette, A. F., Desnos-Ollivier, M., Ni, M., Dromer, F., & Heitman, J. (2012). Genetic Diversity and Genomic Plasticity of *Cryptococcus neoformans* AD Hybrid Strains. *G3:Genes|Genomes|Genetics, 2(1),* 83–97. 10.1534/g3.111.001255

Lipinski, K. A., Puchta, O., Surendranath, V., Kudla, M., & Golik, P. (2011). Revisiting the yeast PPR proteins--application of an Iterative Hidden Markov Model algorithm reveals new members of the rapidly evolving family. *Molecular Biology and Evolution, 28(10),* 2935–2948. 10.1093/molbev/msr120

Lippman, Z. B., & Zamir, D. (2007). Heterosis: revisiting the magic. *Trends in Genetics, 23(2),* 60–66). 10.1016/j.tig.2006.12.006

Liti, G., Barton, D. B. H., & Louis, E. J. (2006). Sequence diversity, reproductive isolation and species concepts in *Saccharomyces*. *Genetics, 174(2),* 839–850. 10.1534/genetics.106.062166

Lockhart, S. R., Messer, S. A., Pfaller, M. A., & Diekema, D. J. (2008). Geographic Distribution and Antifungal Susceptibility of the Newly Described Species *Candida orthopsilosis* and *Candida metapsilosis* in Comparison to the Closely Related Species *Candida parapsilosis*. *Journal of Clinical Microbiology, 46(8),* 2659–2664. 10.1128/jcm.00803-08

Lodder, J., & Kreger-van Rij, N. J. W. (1952). The Yeasts. A Taxonomic Study, 1st Edn. Amsterdam: North-Holland Publishing Company, 671.

Logue, M. E., Wong, S., Wolfe, K. H., & Butler, G. (2005). A Genome Sequence Survey Shows that the Pathogenic Yeast *Candida parapsilosis* Has a Defective *MTLa1* Allele at Its Mating Type Locus. *Eukaryotic Cell, 4(6),* 1009–1017. 10.1128/ec.4.6.1009-1017.2005

Lopes Colombo, A., Azevedo Melo, A. S., Crespo Rosas, R. F., Salomão, R., Briones, M., Hollis, R. J., Messer, S. A., & Pfaller, M. A. (2003). Outbreak of *Candida rugosa* candidemia: an emerging pathogen that may be refractory to amphotericin B therapy. *Diagnostic Microbiology and Infectious Disease, 46(4),* 253–257. 10.1016/s0732-8893(03)00079-8

Louis, V. L., Despons, L., Friedrich, A., Martin, T., Durrens, P., Casarégola, S., Neuvéglise, C., Fairhead, C., Marck, C., Cruz, J. A., Straub, M.-L., Kugler, V., Sacerdot, C., Uzunov, Z., Thierry, A., Weiss, S., Bleykasten, C., De Montigny, J., Jacques, N., … Souciet, J.-L. (2012). *Pichia sorbitophila*, an Interspecies Yeast Hybrid, Reveals Early Steps of Genome Resolution After Polyploidization. *G3:Genes|Genomes|Genetics, 2(2)*, 299–311. 10.1534/g3.111.000745

Lunt, D. H., Kumar, S., Koutsovoulos, G., & Blaxter, M. L. (2014). The complex hybrid origins of the root knot nematodes revealed through comparative genomics. *PeerJ, 2,* e356. 10.7717/peerj.356

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., … Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience, 1(1),* 18. 10.1186/2047-217X-1-18

Magee, B. B., & Magee, P. T. (2000). Induction of mating in *Candida albicans* by construction of *MTLa* and *MTLalpha* strains. *Science, 289(5477),* 310–313. 10.1126/science.289.5477.310

Majoros, L., Kardos, G., Belák, A., Maráz, A., Asztalos, L., Csánky, E., et al. (2003). Restriction enzyme analysis of ribosomal DNA shows that *Candida inconspicua* clinical isolates can be misidentified as *Candida norvegensis* with traditional diagnostic procedures. *Journal of Clinical Microbiology. 41,* 5250–5253. 10.1128/jcm.41.11.5250-5253.2003

Majoros, L., Kardos, G., Szabó, B., Kovács, M., and Maráz, A. (2005). Fluconazole susceptibility testing of *Candida inconspicua* clinical isolates: comparison of four methods. *Journal of Antimicrobial Chemotherapy 55*, 275–276. 10.1093/jac/dkh539

Mallet, J. (2007). Hybrid speciation. *Nature, 446(7133),* 279–283. 10.1038/nature05706

Mallet, J., Beltrán, M., Neukirchen, W., & Linares, M. (2007). Natural hybridization in heliconiine butterflies: the species boundary as a continuum. *BMC Evolutionary Biology,* 7, 28. 10.1186/1471-2148-7-28

Man in ‘t Veld, W. A., de Cock, A. W. A. M., & Summerbell, R. C. (2007). Natural hybrids of resident and introduced Phytophthora species proliferating on multiple new hosts. *European Journal of Plant Pathology, 117(1),* 25-33. 10.1007/s10658-006-9065-9

Mapleson, D., Accinelli, G. G., Kettleborough, G., Wright, J., & Clavijo, B. J. (2017). KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics, 33(4),* 574-576. 10.1093/bioinformatics/btw663

Marcet-Houben, M., & Gabaldón, T. (2015). Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. *PLoS Biology, 13(8),* e1002220. 10.1371/journal.pbio.1002220

Marsit, S., & Dequin, S. (2015). Diversity and adaptive evolution of *Saccharomyces* wine yeast: a review. *FEMS Yeast Research, 15(7),* fov067. 10.1093/femsyr/fov067

Martin, M., Patterson, M., Garg, S., O Fischer, S., Pisanti, N., Klau, G. W., Schöenhuth, A., & Marschall, T. (2016). WhatsHap: fast and accurate read-based phasing. *bioRxiv*. 10.1101/085050

Masuelli, R. W., Camadro, E. L., Erazzú, L. E., Bedogni, M. C., & Marfil, C. F. (2009). Homoploid hybridization in the origin and evolution of wild diploid potato species. *Plant Systematics and Evolution, 277(3-4),* 143–151. 10.1007/s00606-008-0116-x

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science, 226(4676),* 792–801. 10.1126/science.15739260

McGrath, C. L., Gout, J.-F., Johri, P., Doak, T. G., & Lynch, M. (2014). Differential retention and divergent resolution of duplicate genes following whole-genome duplication. *Genome Research, 24(10),* 1665–1675. 10.1101/gr.173740.114

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research, 20(9),* 1297–1303. 10.1101/gr.107524.110

Melo, A. S., Bizerra, F. C., Freymüller, E., Arthington-Skaggs, B. A., & Colombo, A. L. (2011). Biofilm production and evaluation of antifungal susceptibility amongst clinical *Candida* spp. isolates, including strains of the *Candida parapsilosis* complex. *Medical Mycology, 49(3),* 253–262. 10.3109/13693786.2010.530032

Mendel, G. (1866). “Versuche über Pflanzenhybriden”. *Verhandlungen des naturforschenden Vereines in Brünn. 4,* 3-47.

Minces, L. R., Ho, K. S., Veldkamp, P. J., & Clancy, C. J. (2009). *Candida rugosa*: a distinctive emerging cause of candidaemia. A case report and review of the literature. *Scandinavian journal of infectious diseases, 41(11-12),* 892–897. 10.3109/00365540903161531

Minervini, F., Montagna, M. T., Spilotros, G., Monaci, L., Santacroce, M. P., & Visconti, A. (2001). Survey on mycoflora of cow and buffalo dairy products from Southern Italy. *International Journal of Food Microbiology, 69,* 141–146. 10.1016/s01681605(01)00583- 9

Mixão,V. & Gabaldón,T. (2018). Hybridization and emergence of virulence in opportunistic human yeast pathogens. *Yeast, 35,* 5–20. 10.1002/yea.3242

Mixão, V., Hansen, A. P., Saus, E., Boekhout, T., Lass-Florl, C., & Gabaldón, T. (2019). Whole-Genome Sequencing of the Opportunistic Yeast Pathogen *Candida inconspicua* Uncovers Its Hybrid Origin. *Frontiers in Genetics, 10,* 383. 10.3389/fgene.2019.00383

Mixão, V., & Gabaldón, T. (2020). Genomic evidence for a hybrid origin of the yeast opportunistic pathogen *Candida albicans. BMC biology, 18(1),* 48. 10.1186/s12915-020-00776-6

Monerawela, C., & Bond, U. (2018). The hybrid genomes of *Saccharomyces pastorianus*: A current perspective. *Yeast , 35(1),* 39–50. 10.1002/yea.3250

Moraes, M. E., Rosa, C. A., & Sene, F. M. (2005). Preliminary notes on yeasts associated with necrotic cactus stems from different localities in Brazil. *Brazilian journal of biology = Revista brasleira de biologia, 65(2),* 299–304. 10.1590/s1519-69842005000200014

Morales, L., & Dujon, B. (2012). Evolutionary role of interspecies hybridization and genetic exchanges in yeasts. *Microbiology and*

Molecular Biology Reviews, 76(4), 721–739. 10.1128/MMBR.00022-12

Morard, M., Benavent-Gil, Y., Ortiz-Tovar, G., Pérez-Través, L., Querol, A., Toft, C., & Barrio, E. (2020). Genome structure reveals the diversity of mating mechanisms in *Saccharomyces cerevisiae* x *Saccharomyces kudriavzevii* hybrids, and the genomic instability that promotes phenotypic diversity. *Microbial genomics, 6(3),* e000333. 10.1099/mgen.0.000333

Moretti, A., Piergili Fioretti, D., Boncio, L., Pasquali, P., & Del Rossi, E. (2000). Isolation of *Candida rugosa* from turkeys. *Journal of veterinary medicine. B, Infectious diseases and veterinary public health, 47(6),* 433–439. 10.1046/j.1439-0450.2000.00367.x

Mostowy, R., Croucher, N. J., Andam, C. P., Corander, J., Hanage, W. P., & Marttinen, P. (2017). Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. *Molecular biology and evolution, 34(5),* 1167–1182. 10.1093/molbev/msx066

Moyes, D. L., Wilson, D., Richardson, J. P., Mogavero, S., Tang, S. X., Wernecke, J., Höfs, S., Gratacap, R. L., Robbins, J., Runglall, M., Murciano, C., Blagojevic, M., Thavaraj, S., Förster, T. M., Hebecker, B., Kasper, L., Vizcay, G., Iancu, S. I., Kichik, N., … Naglik, J. R. (2016). Candidalysin is a fungal peptide toxin critical for mucosal infection. *Nature, 532(7597),* 64–68. 10.1038/nature17625

Muller, H. J. (1942). Isolating mechanisms, evolution, and temperature. In T. Dobzhansky (Ed.), Biological Symposia: A Series of Volumes Devoted to Current Symposia in the Field of Biology (Vol. 6, pp. 71–125). Jaques Cattell Press.

Muller, H., Hennequin, C., Gallaud, J., Dujon, B., & Fairhead, C. (2008). The asexual yeast *Candida glabrata* maintains distinct a and alpha haploid mating types. *Eukaryotic Cell, 7(5),* 848–858. 10.1128/EC.00456-07

Muzzey, D., Schwartz, K., Weissman, J. S., & Sherlock, G. (2013). Assembly of a phased diploid *Candida albicans* genome facilitates allele-specific measurements and provides a simple model for repeat

and indel structure. *Genome Biology, 14(9),* R97. 10.1186/gb-2013-14-9-r97

Naglik, J. R., Challacombe, S. J., & Hube, B. (2003). *Candida albicans* secreted aspartyl proteinases in virulence and pathogenesis. *Microbiology and molecular biology reviews*, *67(3),* 400–428. https://doi.org/10.1128/mmbr.67.3.400-428.2003

Naglik, J. R., Gaffen, S. L., & Hube, B. (2019). Candidalysin: discovery and function in *Candida albicans* infections. *Current Opinion in Microbiology, 52,* 100–109. 10.1016/j.mib.2019.06.002

NCBI Resource Coordinators (2015). Database resources of the National Center for Biotechnology Information. *Nucleic acids research, 43(Database issue),* D6–D17. 10.1093/nar/gku1130

Neafsey, D. E., Barker, B. M., Sharpton, T. J., Stajich, J. E., Park, D. J., Whiston, E., -Y. Hung, C., McMahan, C., White, J., Sykes, S., Heiman, D., Young, S., Zeng, Q., Abouelleil, A., Aftuck, L., Bessette, D., Brown, A., FitzGerald, M., Lui, A., … Rounsley, S. D. (2010). Population genomic sequencing of *Coccidioides* fungi reveals recent hybridization and transposon control. *Genome Research, 20(7),* 938–946. 10.1101/gr.103911.109

Nedret Koç, A., Kocagöz, S., Erdem, F., & Gündüz, Z. (2002). Outbreak of nosocomial fungemia caused by *Candida glabrata*. *Mycoses, 45(11-12),* 470–475. 10.1046/j.1439-0507.2002.00805.x

Nobile, C. J., & Mitchell, A. P. (2006). Genetics and genomics of *Candida albicans* biofilm formation. *Cellular Microbiology, 8(9),* 1382–1391. 10.1111/j.1462-5822.2006.00761.x

Nucci, M., & Anaissie, E. (2007). *Fusarium* Infections in Immunocompromised Patients. *Clinical Microbiology Reviews, 20(4),* 695–704. 10.1128/cmr.00014-07

O'Brien, C. E., McCarthy, C., Walshe, A. E., Shaw, D. R., Sumski, D. A., Krassowski, T., Fitzpatrick, D. A., & Butler, G. (2018). Genome analysis of the yeast *Diutina catenulata*, a member of the Debaryomycetaceae/Metschnikowiaceae (CTG-Ser) clade. *PloS one, 13(6),* e0198957. 10.1371/journal.pone.0198957

O'Connell, J., Schulz-Trieglaff, O., Carlson, E., Hims, M. M., Gormley, N. A., & Cox, A. J. (2015). NxTrim: optimized trimming of Illumina mate pair reads. *Bioinformatics, 31(12),* 2035–2037. 10.1093/bioinformatics/btv057

O'Donnell, K., Sutton, D. A., Fothergill, A., McCarthy, D., Rinaldi, M. G., Brandt, M. E., Zhang, N., & Geiser, D. M. (2008). Molecular Phylogenetic Diversity, Multilocus Haplotype Nomenclature, and *In Vitro* Antifungal Resistance within the *Fusarium solani* Species Complex. *Journal of Clinical Microbiology, 46(8),* 2477–2490. 10.1128/jcm.02371-07

Odio, C. D., Marciano, B. E., Galgiani, J. N., & Holland, S. M. (2017). Risk Factors for Disseminated Coccidioidomycosis, United States. *Emerging Infectious Diseases, 23(2),* 308-311. 10.3201/eid2302.160505

Ortiz-Merino, R. A., Kuanyshev, N., Braun-Galleani, S., Byrne, K. P., Porro, D., Branduardi, P., & Wolfe, K. H. (2017). Evolutionary restoration of fertility in an interspecies hybrid yeast, by whole-genome duplication after a failed mating-type switch. *PLoS Biology, 15(5),* e2002128. 10.1371/journal.pbio.2002128

Ottenburghs, J. (2019). Multispecies hybridization in birds. *Avian Research, 10(1),* 229.

Pammi, M., Holland, L., Butler, G., Gacser, A., & Bliss, J. M. (2013). *Candida parapsilosis* is a significant neonatal pathogen: a systematic review and meta-analysis. *The Pediatric Infectious Disease Journal, 32(5),* e206–e216. 10.1097/INF.0b013e3182863a1c

Pan, W., Zhao, Y., Xu, Y., & Zhou, F. (2014). WinHAP2: an extremely fast haplotype phasing program for long genotype sequences. *BMC bioinformatics, 15,* 164. 10.1186/1471-2105-15-164

Papon, N., Courdavault, V., Clastre, M., & Bennett, R. J. (2013). Emerging and emerged pathogenic *Candida* species: beyond the

Candida albicans paradigm. *PLoS Pathogens, 9(9),* e1003550. 10.1371/journal.ppat.1003550

Pappas, P. G. (2013). Cryptococcal infections in non-HIV-infected patients. *Transactions of the American Clinical and Climatological Association, 124,* 61–79.

Park, B. J., Wannemuehler, K. A., Marston, B. J., Govender, N., Pappas, P. G., & Chiller, T. M. (2009). Estimation of the current global burden of cryptococcal meningitis among persons living with HIV/AIDS. *AIDS, 23(4),* 525–530. 10.1097/qad.0b013e328322ffac

Payseur, B. A., & Rieseberg, L. H. (2016). A genomic perspective on hybridization and speciation. *Molecular Ecology, 25(11),* 2337–2360. 10.1111/mec.13557

Pegueroles, C., Mixão, V., Carreté, L., Molina, M., & Gabaldón, T. (2020). HaploTypo: a variant-calling pipeline for phased genomes. *Bioinformatics , 36(8),* 2569–2571. 10.1093/bioinformatics/btz933

Pérez-Hansen, A., Lass-Flörl, C., Lackner, M., & Rare Yeast Study Group. (2019). Antifungal susceptibility profiles of rare ascomycetous yeasts. *Journal of Antimicrobial Chemotherapy, 74,* 2649–2656. 10.1093/ jac/dkz231

Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K., Llored, A., Cruaud, C., Labadie, K., Aury, J.-M., Istace, B., Lebrigand, K., Barbry, P., Engelen, S., Lemainque, A., Wincker, P., … Schacherer, J. (2018). Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature, 556(7701),* 339–344. 10.1038/s41586-018-0030-5

Petronczki, M., Siomos, M. F., & Nasmyth, K. (2003). Un Ménage à Quatre. *Cell, 112(4),* 423–440. 10.1016/s0092-8674(03)00083-7

Pfaller, M. A., & Diekema, D. J. (2004). Rare and emerging opportunistic fungal pathogens: concern for resistance beyond *Candida albicans* and *Aspergillus fumigatus*. *Journal of Clinical Microbiology, 42(10),* 4419–4431. 10.1128/JCM.42.10.4419-4431.2004

Pfaller, M. A., Diekema, D. J., Colombo, A. L., Kibbler, C., Ng, K. P., Gibbs, D. L., & Newell, V. A. (2006). *Candida rugosa*, an emerging fungal pathogen with resistance to azoles: geographic and temporal trends from the ARTEMIS DISK antifungal surveillance program. *Journal of clinical microbiology, 44(10),* 3578–3582. 10.1128/JCM.00863-06

Pfaller, M. A., & Diekema, D. J. (2007). Epidemiology of invasive candidiasis: a persistent public health problem. *Clinical Microbiology Reviews, 20(1),* 133–163. 10.1128/CMR.00029-06

Pfaller, M. A., Diekema, D. J., Gibbs, D. L., Newell, V. A., Barton, R., Bijie, H., Bille, J., Chang, S.-C., da Luz Martins, M., Duse, A., Dzierzanowska, D., Ellis, D., Finquelievich, J., Gould, I., Gur, D., Hoosen, A., Lee, K., Mallatova, N., Mallie, M., … Global Antifungal Surveillance Group. (2010a). Geographic variation in the frequency of isolation and fluconazole and voriconazole susceptibilities of *Candida glabrata*: an assessment from the ARTEMIS DISK Global Antifungal Surveillance Program. *Diagnostic Microbiology and Infectious Disease, 67(2),* 162–171. 10.1016/j.diagmicrobio.2010.01.002

Pfaller, M. A., Diekema, D. J., Gibbs, D. L., Newell, V. A., Ellis, D., Tullio, V., Rodloff, A., Fu, W., Ling, T. A., & Global Antifungal Surveillance Group. (2010b). Results from the ARTEMIS DISK Global Antifungal Surveillance Study, 1997 to 2007: a 10.5-year analysis of susceptibilities of *Candida* Species to fluconazole and voriconazole as determined by CLSI standardized disk diffusion. *Journal of Clinical Microbiology, 48(4),* 1366–1377. 10.1128/JCM.02117-09

Pfaller, M., Neofytos, D., Diekema, D., Azie, N., Meier-Kriesche, H. U., Quan, S. P., & Horn, D. (2012). Epidemiology and outcomes of candidemia in 3648 patients: data from the Prospective Antifungal Therapy (PATH Alliance®) registry, 2004-2008. *Diagnostic microbiology and infectious disease, 74(4),* 323–331. 10.1016/j.diagmicrobio.2012.10.003

Pinhati, H. M. S., Casulari, L. A., Souza, A. C. R., Siqueira, R. A., Damasceno, C. M. G., & Colombo, A. L. (2016). Outbreak of candidemia caused by fluconazole resistant *Candida parapsilosis*

strains in an intensive care unit. *BMC Infectious Diseases, 16(1),* 433. 10.1186/s12879-016-1767-9

Porman, A. M., Alby, K., Hirakawa, M. P., & Bennett, R. J. (2011). Discovery of a phenotypic switch regulating sexual mating in the opportunistic fungal pathogen *Candida tropicalis*. *Proceedings of the National Academy of Sciences of the United States of America, 108(52),* 21158–21163. 10.1073/pnas.1112076109

Pretorius, I. S., & Bauer, F. F. (2002). Meeting the consumer challenge through genetically customized wine-yeast strains. *Trends in Biotechnology, 20(10),* 426–432. 10.1016/s0167-7799(02)02049-8

Pryszcz, L. P., Németh, T., Gácser, A., & Gabaldón, T. (2013). Unexpected genomic variability in clinical and environmental strains of the pathogenic yeast *Candida parapsilosis*. *Genome Biology and Evolution, 5(12),* 2382–2392. 10.1093/gbe/evt185

Pryszcz, L. P., Németh, T., Gácser, A., & Gabaldón, T. (2014). Genome comparison of *Candida orthopsilosis* clinical strains reveals the existence of hybrids between two distinct subspecies. *Genome Biology and Evolution, 6(5),* 1069–1078. 10.1093/gbe/evu082

Pryszcz, L. P., Németh, T., Saus, E., Ksiezopolska, E., Hegedűsová, E., Nosek, J., Wolfe, K. H., Gacser, A., & Gabaldón, T. (2015). The Genomic Aftermath of Hybridization in the Opportunistic Pathogen *Candida metapsilosis. PLoS Genetics, 11(10),* e1005626. 10.1371/journal.pgen.1005626

Pryszcz, L. P., & Gabaldón, T. (2016). Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research, 44(12),* e113. 10.1093/nar/gkw294

Pujol, C., Daniels, K. J., Lockhart, S. R., Srikantha, T., Radke, J. B., Geiger, J., & Soll, D. R. (2004). The closely related species *Candida albicans* and *Candida dubliniensis* can mate. *Eukaryotic Cell, 3(4),* 1015–1027. 10.1128/EC.3.4.1015-1027.2004

Qi, L., Fan, W., Xia, X., Yao, L., Liu, L., Zhao, H., Kong, X., & Liu, J. (2018). Nosocomial outbreak of *Candida parapsilosis* sensu stricto fungaemia in a neonatal intensive care unit in China. *The Journal of Hospital Infection, 100(4),* e246–e252. 10.1016/j.jhin.2018.06.009

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics , 26(6),* 841–842. 10.1093/bioinformatics/btq033

Reedy, J. L., Floyd, A. M., & Heitman, J. (2009). Mechanistic plasticity of sexual reproduction and meiosis in the *Candida* pathogenic species complex. *Current biology, 19(11),* 891–899. 10.1016/j.cub.2009.04.058

Rhodes, J., Desjardins, C. A., Sykes, S. M., Beale, M. A., Vanhove, M., Sakthikumar, S., Chen, Y., Gujja, S., Saif, S., Chowdhary, A., Lawson, D. J., Ponzio, V., Colombo, A. L., Meyer, W., Engelthaler, D. M., Hagen, F., Illnait-Zaragozi, M. T., Alanio, A., Vreulink, J. M., Heitman, J., … Cuomo, C. A. (2017). Tracing Genetic Exchange and Biogeography of *Cryptococcus neoformans* var. grubii at the Global Population Level. *Genetics, 207(1),* 327–346. 10.1534/genetics.117.203836

Riccombeni, A., Vidanes, G., Proux-Wéra, E., Wolfe, K. H., & Butler, G. (2012). Sequence and analysis of the genome of the pathogenic yeast *Candida orthopsilosis. PloS One, 7(4),* e35750. 10.1371/journal.pone.0035750

Robinson R. (2008). Birds do it, bees do it, but Candida albicans does it differently. *PLoS biology, 6(5),* e121. 10.1371/journal.pbio.0060121

Rogers, D. W., McConnell, E., Ono, J., & Greig, D. (2018). Spore-autonomous fluorescent protein expression identifies meiotic chromosome mis-segregation as the principal cause of hybrid sterility in yeast. *PLoS biology, 16*(*11*), e2005066. 10.1371/journal.pbio.2005066

Romanel, A., Lago, S., Prandi, D., Sboner, A., & Demichelis, F. (2015). ASEQ: fast allele-specific studies from next-generation

sequencing data. *BMC medical genomics*, *8*, 9. 10.1186/s12920-015-0084-2

Romeo, O., Tietz, H.-J., & Criseo, G. (2013). Candida africana: Is It a Fungal Pathogen? *Current Fungal Infection Reports, 7(3),* 192–197. 10.1007/sl12281-013-0142-1

Ropars, J., Maufrais, C., Diogo, D., Marcet-Houben, M., Perin, A., Sertour, N., Mosca, K., Permal, E., Laval, G., Bouchier, C., Ma, L., Schwartz, K., Voelz, K., May, R. C., Poulain, J., Battail, C., Wincker, P., Borman, A. M., Chowdhary, A., Fan, S., … d'Enfert, C. (2018). Gene flow contributes to diversification of the major fungal pathogen *Candida albicans*. *Nature communications*, *9(1),* 2253. 10.1038/s41467-018-04787-4

Rosenberg S. M. (2011). Stress-induced loss of heterozygosity in *Candida*: a possible missing link in the ability to evolve. *mBio*, *2(5),* e00200-11. 10.1128/mBio.00200-11

Ruan, J., & Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature methods*, *17*(*2*), 155–158. 10.1038/s41592-019-0669-3

Runemark, A., Vallejo-Marin, M., & Meier, J. I. (2019). Eukaryote hybrid genomes. *PLoS genetics*, *15(11),* e1008404. 10.1371/journal.pgen.1008404

Sabino, R., Sampaio, P., Rosado, L., Videira, Z., Grenouillet, F., & Pais, C. (2015). Analysis of clinical and environmental *Candida parapsilosis* isolates by microsatellite genotyping--a tool for hospital infection surveillance. *Clinical microbiology and infection, 21(10),* 954.e1–954.e9548. 10.1016/j.cmi.2015.06.001

Sabino, R., Veríssimo, C., Pereira, Á. A., & Antunes, F. (2020). *Candida auris*, an Agent of Hospital-Associated Outbreaks: Which Challenging Issues Do We Need to Have in Mind?. *Microorganisms*, *8(2),* 181. 10.3390/microorganisms8020181

Safonova, Y., Bankevich, A., & Pevzner, P. A. (2015). dipSPAdes: Assembler for Highly Polymorphic Diploid Genomes. *Journal of*

computational biology: a journal of computational molecular cell biology, *22(6)*, 528–545. 10.1089/cmb.2014.0153

Saha, K., Sit, N. K., Maji, A., & Jash, D. (2013). Recovery of fluconazole sensitive *Candida ciferrii* in a diabetic chronic obstructive pulmonary disease patient presenting with pneumonia. *Lung India: official organ of Indian Chest Society*, *30(4)*, 338–340. 10.4103/0970-2113.120614

Sai, S., Holland, L. M., McGee, C. F., Lynch, D. B., & Butler, G. (2011). Evolution of mating within the *Candida parapsilosis* species group. *Eukaryotic cell*, *10(4)*, 578–587. 10.1128/EC.00276-10

Samarasinghe, H., You, M., Jenkinson, T. S., Xu, J., & James, T. Y. (2020). Hybridization Facilitates Adaptive Evolution in Two Major Fungal Pathogens. *Genes*, *11(1)*, 101. 10.3390/genes11010101

Sandven, P., Nilsen, K., Digranes, A., Tjade, T., & Lassen, J. (1997). *Candida norvegensis*: a fluconazole-resistant species. *Antimicrobial agents and chemotherapy*, *41(6)*, 1375–1376.

Santos, M. A., Gomes, A. C., Santos, M. C., Carreto, L. C., & Moura, G. R. (2011). The genetic code of the fungal CTG clade. *Comptes rendus biologies*, *334(8-9)*, 607–611. 10.1016/j.crvi.2011.05.008

Sardi, J., Scorzoni, L., Bernardi, T., Fusco-Almeida, A. M., & Mendes Giannini, M. (2013). *Candida* species: current epidemiology, pathogenicity, biofilm formation, natural antifungal products and new therapeutic options. *Journal of medical microbiology*, *62(1)*, 10–24. 10.1099/jmm.0.045054-0

Scaccabarozzi, L., Locatelli, C., Pisoni, G., Manarolla, G., Casula, A., Bronzo, V., & Moroni, P. (2011). Short communication: Epidemiology and genotyping of *Candida rugosa* strains responsible for persistent intramammary infections in dairy cows. *Journal of dairy science*, *94(9)*, 4574–4577. 10.3168/jds.2011-4294

Schelenz, S., Hagen, F., Rhodes, J. L., Abdolrasouli, A., Chowdhary, A., Hall, A., Ryan, L., Shackleton, J., Trimlett, R., Meis, J. F., Armstrong-James, D., & Fisher, M. C. (2016). First hospital outbreak of the globally emerging *Candida auris* in a European

hospital. *Antimicrobial resistance and infection control*, *5*, 35. 10.1186/s13756-016-0132-5

Schröder, M. S., Martinez de San Vicente, K., Prandini, T. H., Hammel, S., Higgins, D. G., Bagagli, E., Wolfe, K. H., & Butler, G. (2016). Multiple Origins of the Pathogenic Yeast *Candida orthopsilosis* by Separate Hybridizations between Two Parental Species. *PLoS genetics*, *12(11)*, e1006404. 10.1371/journal.pgen.1006404

Seervai, R. N., Jones, S. K., Jr, Hirakawa, M. P., Porman, A. M., & Bennett, R. J. (2013). Parasexuality and ploidy change in *Candida tropicalis*. *Eukaryotic cell*, *12(12)*, 1629–1640. 10.1128/EC.00128-13

Session, A. M., Uno, Y., Kwon, T., Chapman, J. A., Toyoda, A., Takahashi, S., Fukui, A., Hikosaka, A., Suzuki, A., Kondo, M., van Heeringen, S. J., Quigley, I., Heinz, S., Ogino, H., Ochi, H., Hellsten, U., Lyons, J. B., Simakov, O., Putnam, N., Stites, J., … Rokhsar, D. S. (2016). Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature*, *538(7625)*, 336–343. 10.1038/nature19840

Shimodaira, H., & Hasegawa, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, *17(12)*, 1246–1247. 10.1093/bioinformatics/17.12.1246

Short, D. P., O'Donnell, K., Thrane, U., Nielsen, K. F., Zhang, N., Juba, J. H., & Geiser, D. M. (2013). Phylogenetic relationships among members of the *Fusarium solani* species complex in human infections and the descriptions of *F. keratoplasticum* sp. nov. and *F. petroliphilum* stat. nov. *Fungal genetics and biology*, *53*, 59–70. 10.1016/j.fgb.2013.01.004

Short, D. P., O'Donnell, K., & Geiser, D. M. (2014). Clonality, recombination, and hybridization in the plumbing-inhabiting human pathogen *Fusarium keratoplasticum* inferred from multilocus sequence typing. *BMC evolutionary biology*, *14*, 91. 10.1186/1471-2148-14-91

Shull, G. H. (1908). The Composition of a Field of Maize. *Journal of Heredity, 4(1),* 296–301. 10.1093/jhered/os-4.1.296

Shull, G. H. (1914). Duplicate genes for capsule-form inBursa bursa-pastoris. *Zeitschrift für Induktive Abstammungs-und Vererbungslehre, 12(1),* 97–149. 10.1007/bf01837282

Simchen G. (2009). Commitment to meiosis: what determines the mode of division in budding yeast? *BioEssays: news and reviews in molecular, cellular and developmental biology, 31(2),* 169–177. 10.1002/bies.200800124

Sionov, E., Lee, H., Chang, Y. C., & Kwon-Chung, K. J. (2010). *Cryptococcus neoformans* overcomes stress of azole drugs by formation of disomy in specific multiple chromosomes. *PLoS pathogens, 6(4),* e1000848. 10.1371/journal.ppat.1000848

Skrzypek, M. S., Binkley, J., Binkley, G., Miyasato, S. R., Simison, M., & Sherlock, G. (2017). The *Candida* Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic acids research, 45(D1),* D592–D596. 10.1093/nar/gkw924

Smith, I. M., Stephan, C., Hogardt, M., Klawe, C., Tintelnot, K., & Rickerts, V. (2015). Cryptococcosis due to *Cryptococcus gattii* in Germany from 2004-2013. *International journal of medical microbiology, 305(7),* 719–723. 10.1016/j.ijmm.2015.08.023

Smukowski Heil, C. S., DeSevo, C. G., Pai, D. A., Tucker, C. M., Hoang, M. L., & Dunham, M. J. (2017). Loss of Heterozygosity Drives Adaptation in Hybrid Yeast. *Molecular biology and evolution, 34(7),* 1596–1612. 10.1093/molbev/msx098

Srikantha, T., Daniels, K. J., Pujol, C., Sahni, N., Yi, S., & Soll, D. R. (2012). Nonsex genes in the mating type locus of *Candida albicans* play roles in a/α biofilm formation, including impermeability and fluconazole resistance. *PLoS pathogens, 8(1),* e1002476. 10.1371/journal.ppat.1002476

Stamatakis A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large

phylogenies. *Bioinformatics*, *30(9)*, 1312–1313. 10.1093/bioinformatics/btu033

Stanke, M., & Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research*, *33(Web Server issue)*, W465–W467. 10.1093/nar/gki458

Stavrou, A. A., Mixão, V., Boekhout, T., & Gabaldón, T. (2018). Misidentification of genome assemblies in public databases: The case of *Naumovozyma dairenensis* and proposal of a protocol to correct misidentifications. *Yeast*, *35(6)*, 425–429. 10.1002/yea.3303

Stukenbrock, E. H., Christiansen, F. B., Hansen, T. T., Dutheil, J. Y., & Schierup, M. H. (2012). Fusion of two divergent fungal individuals led to the recent emergence of a unique widespread pathogen species. *Proceedings of the National Academy of Sciences of the United States of America*, *109(27)*, 10954–10959. 10.1073/pnas.1201403109

Stukenbrock E. H. (2016). The Role of Hybridization in the Evolution and Emergence of New Fungal Plant Pathogens. *Phytopathology*, *106(2)*, 104–112. 10.1094/PHYTO-08-15-0184-RVW

Stupar, R. M., & Springer, N. M. (2006). Cis-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F1 hybrid. *Genetics*, *173(4)*, 2199–2210. 10.1534/genetics.106.060699

Sugita, T., Takeo, K., Ohkusu, M., Virtudazo, E., Takashima, M., Asako, E., Ohshima, F., Harada, S., Yanaka, C., Nishikawa, A., Majoros, L., & Sipiczki, M. (2004). Fluconazole-resistant pathogens *Candida inconspicua* and *C. norvegensis*: DNA sequence diversity of the rRNA intergenic spacer region, antifungal drug susceptibility, and extracellular enzyme production. *Microbiology and immunology*, *48(10)*, 761–766. 10.1111/j.1348-0421.2004.tb03602.x

Sung, P., & Klein, H. (2006). Mechanism of homologous recombination: mediators and helicases take on regulatory

functions. *Nature reviews. Molecular cell biology*, *7(10)*, 739–750. 10.1038/nrm2008

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one*, *6(7)*, e21800. 10.1371/journal.pone.0021800

Suzzi, G., Schirone, M., Martuscelli, M., Gatti, M., Fornasari, M. E., & Neviani, E. (2003). Yeasts associated with Manteca. *FEMS yeast research*, *3(2)*, 159–166. 10.1016/S1567-1356(02)00183-6

Swanson-Wagner, R. A., Jia, Y., DeCook, R., Borsuk, L. A., Nettleton, D., & Schnable, P. S. (2006). All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. *Proceedings of the National Academy of Sciences of the United States of America*, *103(18)*, 6805–6810. 10.1073/pnas.0510430103

Symington, L. S., Rothstein, R., & Lisby, M. (2014). Mechanisms and regulation of mitotic recombination in *Saccharomyces cerevisiae*. *Genetics*, *198(3)*, 795–835. 10.1534/genetics.114.166140

Tavanti, A., Davidson, A. D., Gow, N. A., Maiden, M. C., & Odds, F. C. (2005). *Candida orthopsilosis* and *Candida metapsilosis* spp. nov. to replace *Candida parapsilosis* groups II and III. *Journal of clinical microbiology*, *43(1)*, 284–292. 10.1128/JCM.43.1.284-292.2005

Theelen, B., Silvestri, M., Guého, E., van Belkum, A., & Boekhout, T. (2001). Identification and typing of Malassezia yeasts using amplified fragment length polymorphism (AFLP), random amplified polymorphic DNA (RAPD) and denaturing gradient gel electrophoresis (DGGE). *FEMS yeast research*, *1(2)*, 79–86. 10.1111/j.1567-1364.2001.tb00018.x

Thiemann, A., Fu, J., Schrag, T. A., Melchinger, A. E., Frisch, M., & Scholten, S. (2010). Correlation between parental transcriptome and field data for the characterization of heterosis in Zea mays L. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, *120(2)*, 401–413. 10.1007/s00122-009-1189-9

Thomaz, D. Y., de Almeida, J. N., Jr, Lima, G., Nunes, M. O., Camargo, C. H., Grenfell, R. C., Benard, G., & Del Negro, G. (2018). An Azole-Resistant *Candida parapsilosis* Outbreak: Clonal Persistence in the Intensive Care Unit of a Brazilian Teaching Hospital. *Frontiers in microbiology*, *9*, 2997. 10.3389/fmicb.2018.02997

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, *14(2)*, 178–192. 10.1093/bib/bbs017

Tietz, H. J., Hopp, M., Schmalreck, A., Sterry, W., & Czaika, V. (2001). *Candida africana* sp. nov., a new human pathogen or a variant of *Candida albicans*? *Mycoses*, *44(11-12)*, 437–445. 10.1046/j.1439-0507.2001.00707.x

Tirosh, I., Reikhav, S., Levy, A. A., & Barkai, N. (2009). A yeast hybrid provides insight into the evolution of gene expression regulation. *Science*, *324(5927)*, 659–662. 10.1126/science.1169766

Trofa, D., Gácser, A., & Nosanchuk, J. D. (2008). *Candida parapsilosis*, an emerging fungal pathogen. *Clinical microbiology reviews*, *21(4)*, 606–625. 10.1128/CMR.00013-08

Turner, S. A., & Butler, G. (2014). The *Candida* pathogenic species complex. *Cold Spring Harbor perspectives in medicine*, *4(9)*, a019778. 10.1101/cshperspect.a019778

Tusso, S., Nieuwenhuis, B., Sedlazeck, F. J., Davey, J. W., Jeffares, D. C., & Wolf, J. (2019). Ancestral Admixture Is the Main Determinant of Global Biodiversity in Fission Yeast. *Molecular biology and evolution*, *36(9)*, 1975–1989. 10.1093/molbev/msz126

UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic acids research*, *47(D1)*, D506–D515. 10.1093/nar/gky1049

Upadhyay, S., Wadhwa, T., & Sarma, S. (2018). A Series of Six Cases of *Candida Ciferrii* Infection in a Tertiary Care Centre of

North India. *Journal of Advances in Medicine and Medical Research, 27(10),* 1-5. 10.9734/jammr/2018/43742

Verspohl, A., Pignedoli, S., & Giudici, P. (2018). The inheritance of mitochondrial DNA in interspecific *Saccharomyces* hybrids and their properties in winemaking. *Yeast (Chichester, England), 35(1),* 173–187. 10.1002/yea.3288

Villanueva-Lozano, H., Treviño-Rangel, R. J., Hernández-Balboa, C. L., González, G. M., & Martínez-Reséndez, M. F. (2016). An unusual case of *Candida ciferrii* fungemia in an immunocompromised patient with Crohn's and *Mycobacterium bovis* disease. *Journal of infection in developing countries, 10(10),* 1156–1158. 10.3855/jidc.8228

Vincent, B. M., Lancaster, A. K., Scherz-Shouval, R., Whitesell, L., & Lindquist, S. (2013). Fitness trade-offs restrict the evolution of resistance to amphotericin B. *PLoS biology, 11(10),* e1001692. 10.1371/journal.pbio.1001692

Viviani, M. A., Cogliati, M., Esposto, M. C., Lemmer, K., Tintelnot, K., Colom Valiente, M. F., Swinne, D., Velegraki, A., Velho, R., & European Confederation of Medical Mycology (ECMM) Cryptococcosis Working Group (2006). Molecular analysis of 311 *Cryptococcus neoformans* isolates from a 30-month ECMM survey of cryptococcosis in Europe. *FEMS yeast research, 6(4),* 614–619. 10.1111/j.1567-1364.2006.00081.x

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS one, 9(11),* e112963. 10.1371/journal.pone.0112963

Wang, Z., Ni, Z., Wu, H., Nie, X., & Sun, Q. (2006). Heterosis in root development and differential gene expression between hybrids and their parental inbreds in wheat (*Triticum aestivum* L.). *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik, 113(7),* 1283–1294. 10.1007/s00122-006-0382-3

Wang, J. M., Bennett, R. J., & Anderson, M. Z. (2018). The Genome of the Human Pathogen *Candida albicans* Is Shaped by Mutation and Cryptic Sexual Recombination. *mBio*, *9(5)*, e01205-18. 10.1128/mBio.01205-18

Warschefsky, E., Penmetsa, R. V., Cook, D. R., & von Wettberg, E. J. (2014). Back to the wilds: tapping evolutionary adaptations for resilient crops through systematic hybridization with crop wild relatives. *American journal of botany*, *101(10)*, 1791–1800. 10.3732/ajb.1400116

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2018). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular biology and evolution*, *35(3)*, 543–548. 10.1093/molbev/msx319

Waterhouse, R. M., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2019). Using BUSCO to Assess Insect Genomic Resources. *Methods in molecular biology*, *1858*, 59–74. 10.1007/978-1-4939-8775-7_6

Weiß, C. L., Pais, M., Cano, L. M., Kamoun, S., & Burbano, H. A. (2018). nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC bioinformatics*, *19(1)*, 122. 10.1186/s12859-018-2128-z

Welch, M. E., & Rieseberg, L. H. (2002a). Habitat divergence between a homoploid hybrid sunflower species, *Helianthus paradoxus* (Asteraceae), and its progenitors. *American journal of botany*, *89(3)*, 472–478. 10.3732/ajb.89.3.472

Welch, M. E., & Rieseberg, L. H. (2002b). Patterns of genetic variation suggest a single, ancient origin for the diploid hybrid species *Helianthus paradoxus*. *Evolution; international journal of organic evolution*, *56(11)*, 2126–2137. 10.1111/j.0014-3820.2002.tb00138.x

Wertheimer, N. B., Stone, N., & Berman, J. (2016). Ploidy dynamics and evolvability in fungi. *Philosophical transactions of the Royal*

Society of London. Series B, Biological sciences, *371(1709)*, 20150461. 10.1098/rstb.2015.0461

Willems, H., Lowes, D. J., Barker, K. S., Palmer, G. E., & Peters, B. M. (2018). Comparative Analysis of the Capacity of the *Candida* Species To Elicit Vaginal Immunopathology. *Infection and immunity*, *86(12)*, e00527-18. 10.1128/IAI.00527-18

Wolfe K. H. (2015). Origin of the Yeast Whole-Genome Duplication. *PLoS biology*, *13(8)*, e1002221. 10.1371/journal.pbio.1002221

Wu, G., Zhao, H., Li, C., Rajapakse, M. P., Wong, W. C., Xu, J., Saunders, C. W., Reeder, N. L., Reilman, R. A., Scheynius, A., Sun, S., Billmyre, B. R., Li, W., Averette, A. F., Mieczkowski, P., Heitman, J., Theelen, B., Schröder, M. S., De Sessions, P. F., Butler, G., … Dawson, T. L., Jr (2015). Genus-Wide Comparative Genomics of *Malassezia* Delineates Its Phylogeny, Physiology, and Niche Adaptation on Human Skin. *PLoS genetics*, *11(11)*, e1005614. 10.1371/journal.pgen.1005614

Xiao, M., Fan, X., Chen, S. C., Wang, H., Sun, Z. Y., Liao, K., Chen, S. L., Yan, Y., Kang, M., Hu, Z. D., Chu, Y. Z., Hu, T. S., Ni, Y. X., Zou, G. L., Kong, F., & Xu, Y. C. (2015). Antifungal susceptibilities of *Candida glabrata* species complex, *Candida krusei*, *Candida parapsilosis* species complex and *Candida tropicalis* causing invasive candidiasis in China: 3 year national surveillance. *The Journal of antimicrobial chemotherapy*, *70(3)*, 802–810. 10.1093/jac/dku460

Xu, J., Luo, G., Vilgalys, R. J., Brandt, M. E., & Mitchell, T. G. (2002). Multiple origins of hybrid strains of *Cryptococcus neoformans* with serotype AD. *Microbiology (Reading, England)*, *148(1)*, 203–212. 10.1099/00221287-148-1-203

Xu, W., Pang, K. L., & Luo, Z. H. (2014). High fungal diversity and abundance recovered in the deep-sea sediments of the Pacific Ocean. *Microbial ecology*, *68(4)*, 688–698. 10.1007/s00248-014-0448-8

Yarrow, D., and Meyer, S. A. (1978). Proposal for amendment of the diagnosis of the genus *Candida* Berkhout nom. cons. *International*

journal of systematic bacteriology, 28, 611–615. 10.1099/00207713-28-4-611

Ye, C., Hill, C. M., Wu, S., Ruan, J., & Ma, Z. S. (2016). DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. *Scientific Reports, 6,* 31900. 10.1038/srep31900

Yin, Y., & Petes, T. D. (2013). Genome-wide high-resolution mapping of UV-induced mitotic recombination events in *Saccharomyces cerevisiae*. *PLoS Genetics, 9(10),* e1003894. 10.1371/journal.pgen.1003894

Zhai, B., Ola, M., Rolling, T., Tosini, N. L., Joshowitz, S., Littmann, E. R., . . . Hohl, T. M. (2020). High-resolution mycobiota analysis reveals dynamic intestinal translocation preceding invasive candidiasis. *Nature Medicine, 26(1),* 59-64. 10.1038/s41591-019-0709-7

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology, 7(1-2),* 203-214. 10.1089/10665270050081478

Zhang, Z., Bendixsen, D. P., Janzen, T., Nolte, A. W., Greig, D., & Stelkens, R. (2020). Recombining Your Way Out of Trouble: The Genetic Architecture of Hybrid Fitness under Environmental Stress. *Molecular Biology and Evolution, 37(1),* 167-182. 10.1093/molbev/msz211

Zhu, Y. O., Sherlock, G., & Petrov, D. A. (2016). Whole Genome Analysis of 132 Clinical *Saccharomyces cerevisiae* Strains Reveals Extensive Ploidy Variation. *G3, Genes|Genomes|Genetics, 6(8),* 2421–2434. 10.1534/g3.116.029397

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics, 29(21),* 2669-2677. 10.1093/bioinformatics/btt476