Comparative transcriptomics of host-pathogen interactions and hybridization in *Candida* pathogens

Hrant Hovhannisyan

TESI DOCTORAL UPF / 2020

Thesis supervisor
Dr. Toni Gabaldón Estevan

Comparative Genomics group
Life Sciences Department
Barcelona Supercomputing Centre

Pompeu Fabra University
Department of Experimental and Health Sciences

upf. Universitat
Pompeu Fabra
Barcelona

BSC Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Acknowledgments

It was a long and exciting journey, which gave me knowledge and experience expanding my mind, and people whom now I can call my good friends. At the end of this journey, I would like to thank everyone involved in it, though of course these words cannot express all my gratitude.

I thank all my dear people from the Comparative Genomics group who made me feel at home from the first days I actually have left it.

I thank all OPATHIANs for being great and bright people, and for making the whole PhD way more fun.

I thank European Commission and EU taxpayers for realizing that science matters.

Toni, thank you for giving me this chance and your constant support along this whole journey!

And Karine, thank you for being beside me despite many things.

Abstract

Human fungal pathogens are a global healthcare problem. *Candida* yeasts, the most widespread opportunistic human fungal pathogens, comprise phylogenetically diverse species, including newly emerged pathogens. How human-*Candida* interactions vary across species, and what processes underlie the emergence of novel *Candida* pathogens are poorly understood questions. Here we addressed these issues from a comparative transcriptomics perspective. We studied the interactions of human epithelial cells with the four main *Candida* pathogens, and found highly specific transcriptome responses in the various yeasts. Human cells initially responded to the four pathogens with a common response, which we here show to be regulated through mitochondria. At the later stages, host cells responded differently to each pathogen, in a way driven by the level of damage elicited by each species. We predicted lncRNAs in these species and assessed their potential role in infection. Further, we designed a pan-*Candida* targeted enrichment kit, and validated its potent ability to enrich fungal transcriptomes from human-derived samples. Hybridization has been proposed to contribute to the emergence of new yeast pathogens, and we here investigated the transcriptional aftermath of hybridization. In hybrid pathogenic strains of *Candida orthopsilosis* we demonstrated that hybridization moderately affects expression levels, but found allele-specifically expressed genes related to virulence, and showed that some of them might be under selection. We explored the aftermath of hybridization in a newly formed hybrid between *Saccharomyces cerevisiae* and *Saccharomyces uvarum*. We confirmed a moderate effect of gene expression, and found that inter-species transcriptional differences are attenuated. Chromatin accessibility patterns were also conserved. Finally, we developed Crossmapper, a bioinformatics tool facilitating experimental planning of multi-species sequencing studies. Altogether, the results of this thesis expand our knowledge on relevant aspects of host-pathogen interactions and yeast evolution.

Resumen

Los hongos patógenos de humanos son un problema de salud global. Las levaduras del género *Candida*, son los hongos patógenos oportunistas más comunes, y comprenden especies filogenéticamente diversas, incluyendo especies patógenas que han aparecido recientemente. De qué manera varían las interacciones entre humanos y las diferentes especies de *Candida*, y qué procesos subyacen a la aparición de nuevos patógenos del género *Candida* son preguntas aún sin respuesta. Aquí abordamos estos problemas desde la perspectiva de la transcriptómica comparativa. Al estudiar las interacciones de las células epiteliales humanas con los cuatro patógenos principales de *Candida*, encontramos que las transcripcionales eran altamente específicas en las diferentes especies de levadura. Por su parte, las células humanas respondieron de una manera común a los diferentes patógenos, y aquí demostramos que esta respuesta está regulada a través de las mitocondrias. En etapas posteriores, las células epiteliales respondieron de manera diferente, según el nivel de daño provocado por cada uno de los patógenos. Hicimos una predicción de lncRNAs en estas especies y evaluamos el posible papel de estas moléculas en la infección. Además, diseñamos un kit de enriquecimiento selectivo pan-*Candida* y validamos su capacidad para enriquecer tránscriptomas fúngicos a partir de muestras derivadas de humanos. Se ha propuesto que la hibridación contribuye a la aparición de nuevos patógenos de levadura, y aquí investigamos las consecuencias transcripcionales de la hibridación. En cepas patógenas híbridas de *C. orthopsilosis*, demostramos que la hibridación afecta moderadamente los niveles de expresión, pero encontramos que genes relacionados con la virulencia tenían una expresión alélica específica, y demostramos que algunos de ellos podrían estar sujetos a selección. Exploramos las consecuencias de la hibridación en un híbrido recién formado entre *S. cerevisiae* y *S. uvarum.* Confirmamos un efecto moderado de la hibridación en la expresión génica, y observamos una atenuación de las diferencias transcripcionales entre especies. Así mismo, encontramos una gran conservación en los patrones de accesibilidad de la cromatina. Finalmente, desarrollamos Crossmapper, una herramienta bioinformática que facilita la planificación experimental de estudios de secuenciación de especies múltiples. En su conjunto, los resultados de esta tesis amplían nuestro conocimiento sobre aspectos relevantes de las interacciones huésped-patógeno y la evolución de las levaduras.

Preface

Opportunistic human fungal pathogens pose a serious threat for global healthcare. In part due to an increasing number of immunocompromised people worldwide, the incidence of fungal infections has been growing in the last two decades (Pfaller and Diekema, 2007; Oren and Paul, 2014). These infections range from superficial rushes, affecting a quarter of the human population, to life-threatening invasive mycoses, which kill 1.5 million people annually (Havlickova, Czaika and Friedrich, 2008; Brown *et al.*, 2012).

Yeasts from the *Candida* clade are among the major opportunistic human fungal pathogens (Guinea, 2014). These fungi normally constitute part of the healthy human microbiome, but in immunocompromised states they can originate severe infections, with mortality rates reaching 60-70% (Flevari *et al.*, 2013; Klingspor *et al.*, 2015). *Candida* pathogens have several peculiarities which make them difficult to deal with. They are phylogenetically diverse, novel pathogenic species are constantly appearing - sometimes through interspecies hybridization - and they often develop resistance to antimycotic drugs (Gabaldón, Naranjo-Ortíz and Marcet-Houben, 2016; Ksiezopolska and Gabaldón, 2018; Mixão and Gabaldón, 2018). Due to these properties, we still lack fast and accurate diagnostics tools and effective approaches for treating candidiasis. Despite the great importance of *Candida* infections for our health, the research on host-pathogen interactions of these yeasts, and fungi in general, has been traditionally somewhat neglected compared to other microorganisms ('Stop neglecting fungi', 2017).

Advances in next-generation sequencing in the previous decade have opened unprecedented possibilities for studying the biology of these fungi. For example, genome sequencing serves as an important tool, which has already allowed to disentangle complex phylogenetic relationships among *Candida* and uncovered hybridization as one of the driving forces underlying the emergence of novel *Candida* pathogens (Pryszcz *et al.*, 2015; Gabaldón, Naranjo-Ortíz and Marcet-Houben, 2016; Mixão *et al.*, 2019).

Transcriptome sequencing by means of RNA-Seq is a powerful technique which dissects the dynamic interactions between host and pathogen at the level of gene expression (Westermann, Barquist and Vogel, 2017). Despite the proven power of this methodology to disentangle the direct and immediate mechanisms of host-pathogen interplay during infection, so far there are only a handful of studies using this technique on *Candida* pathogens. Thus, there are still numerous fundamental and technical issues

in the context of human-*Candida* interactions that can be effectively addressed using transcriptomics.

In this thesis, we applied large scale comparative transcriptomics and bioinformatics to investigate central basic and methodological aspects of host-pathogen interactions and emergence of virulence in *Candida* pathogens. In particular, we investigated the transcriptomic basis of host-pathogen interactions *in vitro* between human epithelium and four major *Candida* species, revealing generalized and specific mechanisms. Apart from the protein coding portion of fungal transcriptomes, we assessed, for the first time, the potential role of long non-coding transcripts in the virulence of these fungi. We additionally made the first steps towards clarifying how hybridization shapes the transcriptomes of yeast hybrids, and whether this shaping has links to the emergence of pathogenesis.

From a methodological perspective, we developed novel experimental approaches and bioinformatics tools that greatly facilitate the study of host-fungus interactions, and which usage can be extended beyond the research topics of this thesis.

Collectively, in the frame of this thesis we exploited the power of large scale comparative transcriptomics for addressing several key aspects of host-microbe interactions and emergence of pathogenicity in *Candida* species. The results of our work significantly extend our understanding of these important questions, and also open new exciting perspectives for future research in the field. Moreover, the tools and the extensive transcriptome sequencing data generated in this project will certainly serve as a rich and powerful resource for the research community.

Thesis Overview

The thesis consists of 10 Chapters, which are briefly outlined below:

Chapter 1 introduces the Kingdom Fungi and the main human fungal pathogens. It further describes currently known molecular mechanisms of host-microbe interactions between *Candida* pathogens and the human host.

Chapter 2 introduces the phenomenon of hybridization and describes current advances of hybridization research in fungi. It also discusses our current understanding of the genomic and transcriptomic aftermaths of hybridization.

Chapter 3 surveys state-of-the-art comparative transcriptomics methodologies and their use to investigate host-pathogens interactions. It also provides the reader with an overview of recent studies performed to elucidate fungal host-pathogen interactions. This chapter has been published as a review in the book series "Current Topics in Microbiology and Immunology".

Chapter 4 presents a large-scale dual transcriptome analysis performed in collaboration with other groups to investigate host-pathogen interactions of the four main *Candida* pathogens (namely *Candida albicans*, *Candida glabrata*, *Candida parapsilosis* and *Candida tropicalis*) with human epithelial cells. This study showed that these four yeasts have remarkably different transcriptional responses towards human epithelia. In contrast, human cells initially had a uniform response to these diverse pathogens, which then diversified depending on the damaging capacity of each of the pathogens. Moreover, we identified a novel mechanism of immune response against fungi, which is governed by host mitochondria. This chapter is currently under revision in "Nature Microbiology" journal.

Chapter 5 describes the identification and characterization of long non-coding RNAs (lncRNAs) in the four main *Candida* species, and the investigation of potential roles of lncRNAs in infection progression.

Chapter 6 presents the validation of a novel targeted enrichment approach to enrich fungal RNA from human-derived samples. We demonstrated that targeted enrichment of RNA by oligonucleotide probes is superior over other enrichment methods, and serves as a powerful tool for performing RNA-Seq studies of host-pathogen interactions *in vivo*.

Chapter 7 represents a study where we investigated the impact of hybridization on the transcriptome profiles of the opportunistic human fungal pathogen *C. orthopsilosis*. We found that the overall transcriptional impact of hybridization is modest and that the hybrid preferentially retains heterozygous alleles which have allele-specific expression. Additionally, we found that some genes that have allele-specific expression were involved in processes related to pathogenicity. This chapter has been published in “mSphere” journal.

Chapter 8 describes a research study of the consequences of hybridization in an artificial yeast hybrid between *S. cerevisiae* and *S. uvarum,* at the levels of transcriptome and chromatin accessibility. We show that the effect of hybridization is buffered and more moderate as compared with an environmental stress, like temperature shock, uncovering the basic patterns of gene expression behaviour of yeasts upon hybridization. The results of this chapter have been published in “Frontiers in Genetics” journal.

Chapter 9 presents Crossmapper - a novel bioinformatics pipeline that allows to design sequencing studies that share a common principle applied in all projects performed in this thesis, i.e. sequencing of samples containing genetic material of several species. Crossmapper is a versatile and highly customizable tool, allowing to calculate read cross-mapping rates before performing the actual sequencing of multi-species samples, thus allowing to plan the sequencing configurations in advance. This software has been published in “Bioinformatics” journal.

Chapter 10 presents a summarizing discussion, followed by **Conclusions** of this thesis.

Finally, the **Appendix** lists all projects which I was involved in during my PhD.

Table of Contents

1 Fungal pathogens

1.1 Fungal diversity and Saccharomycotina yeasts

The Kingdom Fungi is extremely diverse. Estimates of the number of fungal species vary dramatically, from the relatively modest 120.000 currently accepted species to as high as 10 million (Hawksworth and Lücking, 2017), with estimates between 2.2 to 3.8 million fungal species reaching broadest consensus (Hawksworth and Lücking, 2017).

The huge diversity of fungi is also reflected in the environmental niches they occupy - fungi are found everywhere, from the bottom of seas (Orsi, Biddle and Edgcomb, 2013) to stratosphere (Wainwright *et al.*, 2003) and everywhere in between (Naranjo-Ortiz and Gabaldón, 2019).

They can grow in a wide spectrum of temperatures ranging from -10 °*C* to 65 °*C* (Baxter and Illston, 1980; Marquez *et al.*, 2007). Lifestyle of fungi also varies substantially (Rodriguez and Redman, 1997; Zeilinger *et al.*, 2016), and includes free living saprotrophic organisms, mutualists, commensals, parasites and pathogens.

Ascomycota is the most studied fungal clade, which constitutes nearly two thirds, or ca 65000 known species (Blackwell, 2011). Saccharomycotina, or budding yeasts, in turn (Fig. 1.1), is the best studied class of Ascomycota (Dujon and Louis, 2017), and likely the most investigated fungal clade.

Saccharomycotina comprises a single class - Saccharomycetes, representing yeasts with naked asci, which do not form fruiting bodies and are able to propagate assexually by budding. This class - commonly known as budding yeasts - harbors several well-known groups of yeasts, including many with industrial and clinical relevance. For example, the genus *Saccharomyces* includes *Saccharomyces cerevisiae* - a work-horse of molecular biology and genetics, which is the first eukaryotic species whose genome was fully sequenced (Goffeau *et al.*, 1996). Other representatives of this genus are industrially important yeasts used in production of beverages, such as *Saccharomyces pastorianu*s, which is an inter-species hybrid between *S. cerevisiae* and *Saccharomyces eubayanus* (Dujon and Louis, 2017).

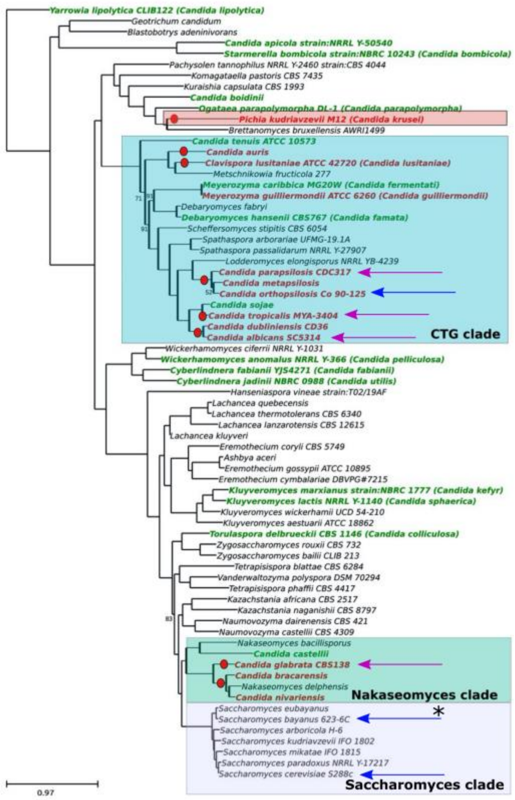

Yarrowia lipolytica CLIB122 (Candida lipolytica)
Geotrichum candidum
Blastobotrys adeninivorans
Candida apicola strain:NRRL Y-50540
Starmerella bombicola strain:NBRC 10243 (Candida bombicola)
Pachysolen tannophilus NRRL Y-2460 strain:CBS 4044
Komagataella pastoris CBS 7435
Kuraishia capsulata CBS 1993
Candida boidinii
Ogataea parapolymorpha DL-1 (Candida parapolymorpha)
Pichia kudriavzevii M12 (Candida krusei)
Brettanomyces bruxellensis AWRI1499
Candida tenuis ATCC 10573
Candida auris
Clavispora lusitaniae ATCC 42720 (Candida lusitaniae)
Metschnikowia fructicola 277
Meyerozyma caribbica MG20W (Candida fermentati)
Meyerozyma guilliermondii ATCC 6260 (Candida guilliermondii)
71
93
Debaryomyces fabryi
Debaryomyces hansenii CBS767 (Candida famata)
91
Scheffersomyces stipitis CBS 6054
Spathaspora arborariae UFMG-19.1A
Spathaspora passalidarum NRRL Y-27907
Lodderomyces elongisporus NRRL YB-4239
Candida parapsilosis CDC317
Candida metapsilosis
52
Candida orthopsilosis Co 90-125
Candida sojae
Candida tropicalis MYA-3404
Candida dubliniensis CD36
Candida albicans SC5314
CTG clade
Wickerhamomyces ciferrii NRRL Y-1031
Wickerhamomyces anomalus NRRL Y-366 (Candida pelliculosa)
Cyberlindnera fabianii YJS4271 (Candida fabianii)
Cyberlindnera jadinii NBRC 0988 (Candida utilis)
Hanseniaspora vineae strain:T02/19AF
Lachancea quebecensis
Lachancea thermotolerans CBS 6340
Lachancea lanzarotensis CBS 12615
Lachancea kluyveri
Eremothecium coryli CBS 5749
Ashbya aceri
Eremothecium gossypii ATCC 10895
Eremothecium cymbalariae DBVPG#7215
Kluyveromyces marxianus strain:NBRC 1777 (Candida kefyr)
Kluyveromyces lactis NRRL Y-1140 (Candida sphaerica)
Kluyveromyces wickerhamii UCD 54-210
Kluyveromyces aestuarii ATCC 18862
Torulaspora delbrueckii CBS 1146 (Candida colliculosa)
Zygosaccharomyces rouxii CBS 732
Zygosaccharomyces bailii CLIB 213
Tetrapisispora blattae CBS 6284
Vanderwaltozyma polyspora DSM 70294
Tetrapisispora phaffii CBS 4417
Kazachstania africana CBS 2517
Kazachstania naganishii CBS 8797
83
Naumovozyma dairenensis CBS 421
Naumovozyma castellii CBS 4309
Nakaseomyces bacillisporus
Candida castellii
Candida glabrata CBS138
Candida bracarensis
Nakaseomyces delphensis
Candida nivariensis
Nakaseomyces clade
Saccharomyces eubayanus
Saccharomyces bayanus 623-6C
*
Saccharomyces arboricola H-6
Saccharomyces kudriavzevii IFO 1802
Saccharomyces mikatae IFO 1815
Saccharomyces paradoxus NRRL Y-17217
Saccharomyces cerevisiae S288c
Saccharomyces clade
0.97

Fig. 1.1. Phylogenetic tree of Saccharomycotina yeasts, zooming into *Candida* and *Saccharomyces* lineages. Adapted from Gabaldón et al. (2016) with modifications. Red-colored species indicate pathogenic *Candida* yeasts while green-colored species indicate non-pathogenic ones. Violet arrows point to the phylogenetic position of the species used for studying host-pathogen interactions in this PhD; Blue arrows indicate species used to study hybridization. ***** - In this project we studied *S. uvarum*

Of note, almost all members of the *Saccharomyces* genus have been reported to form hybrids between each other in both natural and laboratory conditions (Morales and Dujon, 2012). Considering that genomes of *Saccharomyces* yeasts are sequenced, the hybrids of this genus represent a suitable model for studying evolutionary implications of hybridization across a wide range of evolutionary distances. For example, *S. cerevisiae x S. uvarum*, which was used as a model for studying the magnitude of transcriptome shock in this thesis (see Chapter 8), is formed by parentals with 20% of gene divergence (Kellis *et al.*, 2003). Saccharomycetes also includes the diverse group of *Candida* yeasts, which are the most widespread opportunistic human fungal pathogens. This group will be described in more detail below.

With the advent of next-generation sequencing, yeasts of Saccharomycotina are being extensively studied at a genomic level, particularly fueled by their importance in industry and clinics (Gabaldón, Naranjo-Ortíz and Marcet-Houben, 2016; Dujon and Louis, 2017; Shen *et al.*, 2018). The evidence accumulated during recent years indicates that budding yeasts possess high genomic plasticity, with the genomes shaped by several processes including small- and large-scale genomic rearrangements, genome duplications and the ability to form viable hybrids. Such genome rearrangements can result in the emergence of adaptive phenotypes, such as, among others, virulence towards humans (Berman, 2016; Wertheimer, Stone and Berman, 2016). Moreover, genomic plasticity also allows these yeasts to form novel species through hybridization, which appears to be frequent in these fungi (Morales and Dujon, 2012).

1.2 *Candida* pathogens

From the vast diversity of fungi, there are lineages which at certain conditions can be infectious towards humans. These organisms constitute a tiny fraction of the total number of fungal species, with ca 300 currently identified human fungal pathogenic species, which number is gradually increasing (Papon *et al.*, 2013). From an evolutionary perspective, these pathogenic species are scattered across the fungal tree of life, and even within a specific clade the genomic divergence between two pathogenic species (e.g. *Candida glabrata* and *Candida albicans*) can be comparable with the distance between human and fish (Dujon *et al.*, 2004). The most common fungal pathogens are those from *Candida*, *Aspergillus* and *Cryptococcus* clades. Other less frequently encountered species belong to the genera *Histoplasma*, *Blastomyces* and *Malassezia*, among others (Sullivan and Moran, 2014). The lifestyle, and thus the way how these fungi are associated with humans, vary. For instance, the most widespread *Candida* pathogen, *C. albicans* is normally a commensal yeast, constituting a part of the normal human microbiome (Mayer, Wilson and Hube, 2013). On the other hand, all *Aspergillus* species are environmental, and usually infect humans upon inhalation of spores. What is common between most fungal pathogens is their opportunistic nature - normally they become infective in a host with attenuated immune response. Importantly, the number of immunocompromised persons or those with weak immune systems have increased dramatically in the past several decades due to numerous reasons, including but not limited to improved healthcare and emergence and/or increased rates of various diseases (AIDS, cancers, Diabetes, etc). Fungal infections are manifested in numerous forms, ranging from superficial skin rashes to invasive life-threatening forms with high mortality rates. Their opportunistic nature notwithstanding, fungal infections (or mycoses) pose a serious threat for global epidemiology. It is, for example, estimated that a quarter of the global population (Havlickova, Czaika and Friedrich, 2008) is affected by different types of skin nosologies of fungal origin. Moreover, almost 75% of women experience vulvo-vaginal fungal infection throughout their lives (Sobel, 2007). Though invasive mycoses are of course less frequent, they have high rates of mortality, reaching in severe cases up to 70%, and annually kill 1.5 million people worldwide.

The most widespread opportunistic human fungal pathogens belong to Saccharomycotina subphylum and belong to the so-called *Candida* “clade”. Hence the name of the caused nosology - candidiasis. The main infecting agents of these lineages include the widely known species *C. albicans*, *C. glabrata, Candida parapsilosis, Candida tropicalis, Candida auris, Candida krusei,* among others (Gabaldón, Naranjo-Ortíz and Marcet-Houben, 2016). These species are yeasts, some can form true-, others pseudohyphae, and can reproduce asexually by budding.

Despite their common genus name, these species are very diverse and spread across the Saccharomycotina tree (Gabaldón, Naranjo-Ortíz and Marcet-Houben, 2016). Their phylogeny and taxonomy have continuously changed during the past years. For example, *C. parapsilosis* was considered as a single species, until it was identified that it actually comprises a species complex of closely related pathogenic yeasts *C. parapsilosis sensu stricto, C. metapsilosis* and *C. orthopsilosis*, with the latter two being hybrid species (Tavanti *et al.*, 2005; Pryszcz *et al.*, 2014, 2015). Moreover, as in other fungi, the nomenclature and species names in *Candida* clade are sometimes ambiguous (Fig. 1.1). For example, *Candida krusei* is also called *Pichia kudriavzevii, Candida kefyr* is called *Kluyveromyces marxianus, Candida marina* - *Cryptococcus marinus* (despite that true *Cryptococcus* species are from Basidiomycota). Another famous example of naming inconsistency is the case of the second most widespread fungal pathogen *C. glabrata*, which is phylogenetically much closer to species of *Saccharomyces* genus than to other infectious *Candida* species. Altogether, these observations highlight that *Candida* does not constitute a genus or even a clade in a phylogenetic sense, but rather a historically retained naming of diverse yeast species.

In contrast, Saccharomycotina has other well-defined phylogenetic clades. For example, the three most prevalent *Candida* species *C. albicans*, *C. parapsilosis and C. tropicalis* are members of the so-called CTG-clade, which species use an alternative genetic code in which the CUG codon encodes serine instead of usual leucine (Santos *et al.*, 2011). Another peculiarity of *Candida* pathogenic species is that they always have close, non-pathogenic relatives. This, together with differences in virulence mechanisms of these species (described below), indicates that the ability to infect humans is a trait that has likely emerged independently in different lineages.

In the context of emergence of virulence, an increasing number of studies shows compelling evidence that numerous *Candida* pathogens, such as *C. orthopsilosis* (Pryszcz *et al.*, 2015; Schröder, de San Vicente, *et al.*, 2016), *C. metapsilosis* (Pryszcz et al., 2015), *C. inconspicua* (Mixão *et al.*, 2019), and even *C. albicans* (Mixão and Gabaldón, 2020) are interspecific hybrids. This highlights that hybridization can potentially promote the emergence of pathogenicity as an adaptation for colonizing the human host. However, the mechanisms underlying the emergence of virulence through hybridization are still poorly understood and require further research (see Chapter 2 for further details).

Today there are over 30 identified human pathogenic species designated with the *Candida* genus name (Papon *et al.*, 2013), however from the epidemiological perspective around 90% of candidiasis cases are caused by the four most prevalent species *C. albicans*, *C. glabrata*, *C. parapsilosis (sensu stricto)* and *C. tropicalis*, generally in that order. These infections are spread worldwide, but the distribution of the four *Candida* species can vary depending on geographical region, except for *C. albicans* which is usually the most prevalent in all regions. For example, while *C. glabrata* is the second most frequent species in many European countries, *C. parapsilosis* has clearly higher incidence in Spain and Brazil compared with *C. glabrata* (Guinea, 2014).

As stated above, the risk group for developing candidiasis is the immunocompromised population, and thus, these infections are particularly serious in nosocomial (i.e. hospital) conditions and in certain categories of patients (neonates, patients with cancers, AIDS or at intensive care and surgery units). Candidiasis has a variety of forms, such as skin, oral, vulvovaginal and bloodstream (candidemia), which have different severity and expected outcome. In invasive candidiasis, mortality rates are high (up to 70%) and depend on multiple factors, such as age, infecting species/strain, the presence of concomitant diseases, and importantly, the time of diagnosis and medical intervention. Diagnostics of *Candida* pathogens is complicated by numerous factors. Conventional diagnostics include microscopy and biochemical approaches, both of which require fungal cell cultures that are time consuming, and sometimes not possible at all (Ellepola and Morrison, 2005; Arendrup *et al.*, 2014). Even when cultures are grown, morphological or biochemical tests are error-prone, leading to misidentifications. Moreover, diagnostics of these pathogens is further complicated by the emergence of novel infecting species (Papon *et al.*, 2013; Mixão and Gabaldón, 2018) and emergence of antimycotic drug resistance (Satoh *et al.*, 2009; Sanglard, 2016; Ksiezopolska and Gabaldón, 2018), which has to be taken into account for further interventions. More recent methods of molecular diagnostics (reviewed in (Consortium OPATHY and Gabaldón, 2019)), including spectroscopy (e.g. MALDI-TOF) and DNA/RNA testing (e.g. different types Polymerase Change Reactions, PCRs) have higher precision, turnaround time and lower cost, but are not yet fully integrated into current healthcare systems. Thus, despite the fact that molecular methods for the identification of yeast pathogens have developed significantly in recent decades and have shown promising results, it will still take time to turn these techniques into routine clinical practices, even in developed countries.

Summarizing, *Candida* pathogens pose an increasingly serious threat for a considerable and growing part of the human population. So far, the efforts of the scientific, medical and, importantly, healthcare policy makers communities have not been sufficient to ensure a deep understanding of these infections and their effective identification, handling and further treatment.

1.3 Human-*Candida* Interactions

Acquiring a fundamental understanding of the complex interactions between the host and the pathogen is key for disentangling the infection mechanisms of any pathogen, which in turn is the main first step to ensure further advancements. In fact, pathogenicity itself, understood as the capacity of a microbe to harm its host, is the net outcome of interactions between the microbe and the infected host (Casadevall and Pirofski, 2003). This is especially relevant in the context of opportunistic pathogens, as their pathogenicity appears when the balance in host-pathogen interactions is shifted from commensal to a pathogenic state. No matter whether this shift is due to the weakening of the host (for instance because of immunocompromised state) or, conversely, due to reinforcement of the pathogen (e.g. caused by dysregulation of microbiome balance due to antibiotic usage), the outcome of host-microbe interactions will result in pathogenicity manifestation.

Theoretical concepts aside, there are molecular mechanisms acting at different layers of biological complexity which underlie host-pathogen interactions. There are some general factors that influence how the microbes, and specifically fungi, associate with the human host. In this context, the primary factor determining the association of the fungus to the human is host temperature. For most fungi, the relatively high body temperature of endothermic mammals is a hostile environment, and thus only a small proportion of an immense fungal diversity is capable of inhabiting the human niche and acting as pathogens. In particular, Robert and Casadevall (Robert and Casadevall, 2009) estimated that every 1°*C* increase between the 30 °*C*–40 °*C* range excludes an additional 6% of studied fungal organisms from being able to inhabit a given environment. This aspect may underlie why fungi can potentially cause mass extinctions in amphibians (Scheele *et al.*, 2019) whereas having relatively milder effects in mammals.

The molecular mechanisms of host-fungus interactions are multifaceted and intricate, especially when considering the diverse biology of fungi. Moreover, if the interactions between human and certain major pathogens such as *C. albicans* are relatively well-studied, the knowledge about how these interactions vary across other infecting species is scarce.

1.3.1 Virulence mechanisms of *Candida* pathogens

As stated above, *C. albicans* is the most well-studied opportunistic fungal pathogen and its pathogenicity mechanism have been the main focus of research for the last two decades. There are two main aspects that make *C. albicans* a successful pathogen - virulence factors and fitness attributes.

The main virulence factors of *C. albicans* include yeast-hyphal transitions (Felk *et al.*, 2002), the formation of biofilms (Chandra *et al.*, 2001), and the expression of adhesins, invasins, and hydrolytic enzymes (Mayer, Wilson and Hube, 2013; Sullivan and Moran, 2014). Additionally, *C. albicans* has fitness attributes, such as rapid environmental adaptation and metabolic flexibility, stress response and nutrient acquisition systems, which add additional layers of complexity to the virulence of this fungus (Mayer, Wilson and Hube, 2013; Sullivan and Moran, 2014).

Morphotype switching

C. albicans can present several morphological types, including yeast, white or opaque morphology, pseudo- and true-hyphal forms, which are switched among them depending on environmental conditions and signals (Felk *et al.*, 2002). Yeast and hyphal forms are the main morphotypes related to pathogenesis, with the yeast form involved mainly in dissemination, and hyphae formation observed at the actual sites of infection and during invasive growth of candidiasis. The formation of hyphae can be induced by different factors, such as changing pH, temperatures, amount of nutrients (Noble, Gianetti and Witchley, 2017). Though hyphae are directly related to pathogenesis, some specific cell-wall associated or secreted proteins are responsible for *C. albicans* capacity to damage host tissues. For instance, *ECE1* gene encodes a protein that is processed into several peptides, of which one, candidalysin, is a toxin (Moyes *et al.*, 2016). When *ECE1* is deleted, the ability of *C. albicans* to form hyphae is not altered, but the capacity of causing cell damage to the host is impaired. Thus, hyphae formation is necessary but not sufficient to deploy the full pathogenic potential of this fungus.

Adhesion, invasion and damage

Adherence and further invasion of host cells are the primary processes acting during *C. albicans* infection, which are mediated by adhesins and invasins, respectively. The initial adherence to different substrates, including host cells, is achieved via adhesins (Sundstrom, 1999; Liu and Filler, 2011). To date, the most studied adhesins of this fungus are so-called agglutinin-like sequence (Als) adhesins, which form a family (Als1–7 and Als9) of glycosylphosphatidylinositol (GPI)-linked cell surface glycoproteins (Hoyer and Cota, 2016). Others include Hwp1, Eap1, Iff4 and Ecm33 proteins. While some adhesins are morphology independent, others, like Als3 and Hwp1, are associated with hyphae and biofilm formation. Adhesins serve as substrates for host cell-surface proteins, such as transglutaminases, and can covalently bind to them linking the fungus to the host cells. Once attached, *C. albicans* starts host cell invasion by two distinct mechanisms - induced endocytosis and active penetration (Wächtler *et al.*, 2012). Induced endocytosis does not depend on fungal activities, while active penetration, as the name states, is a fungus-dependent process and achieved by actively formed hyphae. Despite being host-driven, endocytosis of the fungus is activated when the host receptors bind to invasins (e.g. Als3 or Ssa1), which are expressed by the fungus (Yang *et al.*, 2014).

From the fungal perspective, the ultimate stage of *C. albicans* infection results in the cellular damage to the host. Until recently, the molecular mechanism of damaging ability of this fungus was elusive. A recent study (Moyes *et al.*, 2016) has identified candidalysin - the first proteolytic fungal peptide toxin. The gene *ECE1* is highly overexpressed (several hundred fold) when the fungus contacts the host cells and starts forming hyphae. After translation of the *ECE1* transcript, the immature Ece1p consists of 271 amino acids (aa), which is further processed by several enzymes such as Kex1p and Kex2p to form the damage-potent peptide of 31 aa. This mature peptide, candidalysin, is then secreted from the hyphae into the engulfed space between the fungus and the host cell, and (at sufficient amounts) forms pores in the host cell membrane, thus leading to cell damage. As mentioned above, deletion of *ECE1* gene was demonstrated to completely restrict the ability of *C. albicans* to cause damage to the host.

Biofilm formation

C. albicans is able to form biofilms on both biotic (in the host) or abiotic (catheters or other medical equipment) surfaces (Chandra *et al.*, 2001; Cavalheiro and Teixeira, 2018). The formation of biofilms is a sequential process, which requires initial adherence to the substrate, microbial expansion, hyphae and extracellular matrix formation and accumulation. On molecular level, biofilm formation is orchestrated by various transcription factors (TF) and downstream effectors, such as Bcr1, Rob1, Hsp90, etc. From the pathogenicity perspective, biofilms have several properties that confer higher resistance against host protective mechanisms or medical interventions. For example, biofilms are more resistant to antimycotic drugs, are generally not affected by neutrophil attacks and do not trigger formation of reactive oxygen species (ROS) in immune cells (Mayer, Wilson and Hube, 2013).

Fitness attributes

Within the host *C. albicans* is exposed to various conditions (Wertheimer, Stone and Berman, 2016). For example, while the environment of the human gut or bloodstream is relatively rich in nutrients (for example, glucose), when the fungus is phagocytosed by a macrophage it is exposed to harsh conditions of limited nutrients and lytic peptides. Thus, flexibility and adaptability for changing environments are crucial for yeast survival. In these varying conditions, *C. albicans* have developed efficient mechanisms for switching metabolic profiles. For example, in niches where carbohydrates are restricted, the fungus can rapidly switch to alternative metabolic pathways to utilize host amino acids and lipids (Lorenz, Bender and Fink, 2004). Similarly, *C. albicans* have multiple mechanisms to respond to stresses imposed by the host, such heat shocks, osmotic and oxidative stresses (Mayer, Wilson and Hube, 2013). For example, the fungus has several types of heat-shock proteins such as Hsp90, Hsp78, Hsp60, etc, which effectively maintain fungal proteins againsts unfolding by high temperatures of the host. On the other hand, *C. albicans* has several quenching proteins, such as Sod1 and Sod5, which can effectively detoxify ROS molecules imposed by phagocytic immune cells.

As stated previously, by far the best studied pathogenicity mechanisms are those of *C. albicans*, and our knowledge about virulence of major non-*C. albicans* species, e.g. *C. glabrata*, *C. parapsilosis* and *C. tropicalis* significantly lags behind. In this context, probably the most relevant focus has to be drawn to *C. glabrata* since phylogenetically it is only distantly related to other major *Candida* yeasts, though it is the second most frequent opportunistic *Candida* pathogen (Gabaldón and Carreté, 2016). Thus, both *C. albicans* and *C. glabrata* overall lead to the same outcome by following presumably different pathogenic paths.

Indeed, as recently reviewed in (Galocha *et al.*, 2019), the major known pathogenicity traits of *C. glabrata* are different from those of *C. albicans*. The initial interactions with the host are done primarily through GPI-linked adhesins encoded by the *EPA* gene family, which was shown to be expanded in *C. glabrata (Gabaldón et al., 2013)*. Yapsin (aspartyl proteases) is the second group of identified pathogenicity-related protein families, which are implicated in cell wall and survival of this fungus within macrophages. Importantly, *C. glabrata* pathogenicity does not depend on morphology switches since it does not form true hyphae, meaning that host invasion is predominantly achieved by induced endocytosis. Unlike *C. albicans*, which actively escapes from the phagocytic macrophage by piercing its membranes with hyphae, *C. glabrata* - which does not form hyphae - is able to persist for a long time and replicate inside macrophages. This leads to high loads of the fungus in the immune cell, which eventually results in lysis of the macrophage. Such kind of adaptation for long-term survival and propagation within harsh macrophage environments indicate that *C. glabrata* has efficient and versatile pathways for stress toleration, which, however, are still poorly understood. Though *C. glabrata* is also able to damage host cells, the mechanisms of damage are still unknown - in contrast with *C. albicans*, this fungus does not seem to produce (or even have) lytic toxins capable of causing damage to the host. Another clear difference between these two fungi are morphologies of their biofilms, while *C. albicans* have relatively thick biofilms with a mixture of yeast-hyphae cells, the biofilm of *C. glabrata* is composed of yeasts-like cells, which makes it thinner but with higher cell density (Rodrigues *et al.*, 2017). Moreover, while for both fungi the formation of biofilms confers higher antibiotic resistance, even in the absence of biofilms, *C. glabrata* has high intrinsic resistance to azoles (Ksiezopolska and Gabaldón, 2018).

Our understanding of virulence mechanisms of *C. parapsilosis* and *C. tropicalis* is even more limited. However, since these two species are phylogenetically close to *C. albicans*, part of the pathogenic arsenal (such as Asl or Hwp proteins) of the latter is also present in the genomes of these two species, though it does not necessarily function in an identical way. For example, while *C. parapsilosis* forms only pseudohyphae (Németh *et al.*, 2013), *C. tropicalis* is capable of forming true hyphae (Jiang *et al.*, 2016), though both of these species cause significantly less damage to the host, suggesting differences in the action of their hyphae-associated hydrolytic enzymes. Moreover, neither of the species was shown to have or express active forms of candidalysin. Corresponding to their morphotypes, biofilms of these species also differ in their densities and thickness.

Despite all the differences in virulence mechanisms between *Candida* species that have been observed so far, our current understanding of the major steps employed by these pathogens follows the same paradigm - first the pathogen adheres to the host tissue, then invades it and subsequently causes damage. Nevertheless, as discussed above, all of the four major *Candida* species follow the same path utilizing different, though in most of the cases poorly understood, mechanisms. This once again reinforces the idea that is proposed solely by observing their phylogenetic relationships - pathogenetic capabilities of these species presumably evolved independently in each of these species.

1.3.2 Host antifungal immunity at a glance

Apart from the intrinsic peculiarities of each of the described pathogens, our understanding of the host biology and how the host interacts with the infecting fungus are key aspects for disentangling the underlying mechanisms of infection. Of note, the host immune system is immensely complex and multifaceted, thus here we will describe the main and general aspects of antifungal immunity, scratching only the tip of the iceberg of immune response complexity. More details can be found in recent, exhaustive reviews (Romani, 2011; Lionakis, Iliev and Hohl, 2017; Salazar and Brown, 2018)

As in the case of fungal pathogens, the host response and its interactions are mainly studied for *C. albicans* using different infection models. The primary system for fungal recognition and further defence against them is innate immunity. Pathogens are first recognized through the detection of so-called Pathogen Associated Molecular Patterns (PAMPs), such as β-glucans or N- and O-linked mannans, by Toll-like (e.g. TLR1, TLR2, TLR4), C-type lectin (e.g. Dectin-1, Dectin2, DC-SIGN) and NOD-like (e.g. NLRP3, NLRC4) pattern recognition receptors (Netea *et al.*, 2008; Salazar and Brown, 2018), which subsequently trigger different defence mechanisms against fungi (reviewed in Moyes et al 2014). To date, neutrophils and macrophages are considered to be the main players against fungal infections, and thus most of the research on host-pathogen interactions has been done in the context of these cell types.

Neutrophils are capable of direct killing and clearing fungi from the organism using several mechanisms, such as phagocytosis with further oxidative killing (Segal, 2005) and formation of neutrophil extracellular traps (NETs) (Mosser and Edwards, 2008). Interestingly, though being one of the key defenders against fungal infections, the role of neutrophils is location-specific, and in some cases, as in vaginal infections, they do not have a clear protective role against fungi (Fidel, 2005).

Macrophages also play important roles against fungal infection. These cells bridge innate and adaptive immune systems and are able to act as both phagocytic and antigen presenting cells. In the presence of interferon gamma, macrophages differentiate into fungicidal M1 subclass which can kill fungi by phagocytosis and subsequent reactive oxygen species (ROS) and NO oxidative stress or by activating Th1 and natural killer cells (reviewed in Moyes et al., 2014).

Although neutrophils and macrophages are crucial against fungal pathogens, recent studies have demonstrated an emerging role of epithelial cells in host-pathogens interaction with fungi (Moyes and Naglik, 2011; Naglik and Moyes, 2011; Naglik *et al.*, 2011). Importantly, it is becoming increasingly more apparent that besides just being the first barrier and anchoring point between the host and pathogens, epithelial cells have sophisticated mechanisms to interact with fungi and cross-talk with surrounding human cells (Moreno-Ruiz *et al.*, 2009; Moyes *et al.*, 2010, 2011; Zhu and Filler, 2010). However, despite the research of host-fungus interplay happening on the epithelial surfaces is gaining momentum, we still lack sufficient knowledge about the host-pathogen interactions at this fundamental site of infection.

2 Hybridization and hybrid pathogens

2.1 Hybridization and the emergence of novel pathogens

Hybridization is a type of reticulate evolution that may result in the emergence of novel species and adaptations to new environments. Inter- and intra-species hybridization has been observed in an extremely wide range of taxa, including fungi, plants and animals.

Natural hybridization is not only widespread, but also frequent. For example, it has been estimated that 40% of vascular plant families trace their origins from hybridization events (Ellstrand, Whitkus and Rieseberg, 1996; Mallet, 2005). Similarly, around 10% of birds (Grant and Grant, 1992) and 6% of European mammals (Mallet, 2005) are also able to hybrise. Nevertheless, the frequency of hybridization is supposedly underestimated. Historically the research of hybridization has been primarily focused on plants, thus other taxa are not investigated to such extent. Moreover, until the advent of advanced sequencing technologies researchers were utilizing methods, such as morphometrics, physiological or biochemical markers, etc., that had limited resolution. That is why, certain lineages such as fungi, which hybrids usually do not possess highlighted phenotypic or physiological differences compared to parentals, largely remained neglected from the mainstream hybridization research. Today, next-generation sequencing methods backed with thorough bioinformatics techniques allowed us to realise that hybridization in fungi is much more frequent than has been unticipated before.

Increasing numbers of fungal hybrids are found among clinically, industrially and environmentally important fungal lineages (Stukenbrock *et al.*, 2012; Charlton *et al.*, 2014; Pryszcz *et al.*, 2014, 2015; Monerawela and Bond, 2018). Some fungal clades, such as *Saccharomyces* (Morales and Dujon, 2012; Borneman and Pretorius, 2015), *Candida* (Hagen *et al.*, 2015; Mixão and Gabaldón, 2018), *Cryptococcus* (Hagen *et al.*, 2015) are particularly prone to hybridizations. Importantly, recent studies have shown that several human fungal pathogens, such as *C. orthopsilosis* (Pryszcz *et al.*, 2014; Schröder, de San Vicente, *et al.*, 2016), *C. metapsilosis* (Pryszcz *et al.*, 2015), *C. inconspicua* (Mixão *et al.*, 2019), *Cr. neoformans, x. Cr. gatii (D'Souza et al., 2011)* and *Malassezia furfur* (Wu *et al.*, 2015) are hybrids. For some of them, both parental species are known (e.g. *Cr. neoformans, x. Cr. gatii*), for others only one of the parentals was identified (e.g. *C. orthopsilosis*), while for the rest neither of the parental species were encountered (e.g. *C. metapsilosis*). This suggests a hypothesis that the non-identified parental species are non-pathogenic environmental fungi inhabiting niches other than human or are outcompeted by their more adapted counterparts.

Either of the cases would underlie that hybridization might provide these species with adaptations for inhabiting and thriving on the human host, which can lead to their pathogenic manifestations towards humans (Mixão and Gabaldón, 2018). A similar phenomenon was also observed in fungal pathogens of plants (Stukenbrock, 2016), as in case of *Zymoseptoria pseudotritici* (Stukenbrock *et al.*, 2012) or *Microbotryum* species (Gladieux *et al.*, 2011). However, the exact mechanisms of how the hybridization contributes to the emergence of pathogenicity or in general novel adaptations in fungi are still poorly understood. To address this, one has to understand the short- and long-term consequences of hybridization on different biological systems of the hybrid, primarily on genome and transcriptome, which in fact will be at the root of developing new adaptations.

A first question to ask is what the necessary steps and prerequisites for two genetically divergent species are to form a viable hybrid organism. First, two parentals must be able to mate, which already constitutes a large barrier. If this step is successful, then the genomes of two diverged organisms will coexist in the same cellular environment. Whether or not, and to which extent the two genomes, and thus all the downstream molecular machineries are compatible with each other will define the viability and fitness of the hybrid. The theoretical concept of incompatibilities between divergent organisms was developed long before our understanding of genomes. Named after the researchers who proposed it, the Bateson–Dobzhansky–Muller model suggests that negative epistatic interactions between genetic material of diverged organisms result in reproductive isolation (Bateson, no date; Dobzhansky, 1934). Thus, after the zygote formation, the hybrid has to cope with incompatibilities between the parental genomes in order to survive. The state where the hybrid experiences these immediate incompatibilities is termed as “genomic shock” (McClintock, 1984), indicating that the divergent counterparts of a newly formed hybrid interfere with each other at different, and in fact not only genomic, levels. How the hybrid will cope with “genomic shock” and what will be the outcomes of this coping eventually will define the properties of the hybrid, including novel adaptations.

There are several ways of overcoming these incompatibilities as discussed below, which in essence define the aftermath of hybridization on different layers of biological organization of the hybrid organism.

2.2 Genomic aftermath of yeast hybridization

Today the consequences of hybridization at the genomic level are arguably the most studied. As mentioned above, hybridization leads to coexistence of two genomes which have evolved independently from each other, and thus, depending on the hybridizing species, accumulated certain levels of genomic divergence. This divergence can vary substantially, particularly in the fungal Kingdom. For example, estimated divergence between parental species of a human pathogen *C. metapsilosis* reaches ~4.5% (Pryszcz *et al.*, 2015), while in an industrially important *S. cerevisiae x S. uvarum* hybrid, the divergence between parentals reaches drastic 38% in intergenic regions and 20% in protein-coding genes (Kellis *et al.*, 2003). The genomic divergence between homeologus (homologous chromosomes from each of the hybrid parentals) chromosomes results in highly unstable genomes upon hybrid formation. There are several possible mechanisms by which the incompatible regions in these genomes are resolved, and in essence these mechanisms overall lead to aneuploidies (Wertheimer, Stone and Berman, 2016; Todd, Forche and Selmecki, 2017) and/or loss of heterozygosity (LOH) (Heil *et al.*, 2017). These mechanisms primarily involve gene conversion (Liu *et al.*, 2018), large-scale deletions (Paun *et al.*, 2007) and duplications (Charron *et al.*, 2019), all of which can be further facilitated by selection to stabilize the hybrid genome. Large scale duplication, extending even to the whole genomes, are of a particular interest in recent years. For example, it has been suggested that ancestors of *Saccharomyces* genus were formed by a hybridization event between distantly related parentals, and subsequently have undergone whole-genome duplication (WGD) to restore their ability to undergo sexual reproduction (Marcet-Houben and Gabaldón, 2015).

Taken together, hybrid organisms employ various strategies to stabilize their highly dissimilar genomes, which eventually is required for their viability. However, despite the fact that genome stabilization is a crucial step towards long-term survival and fitness, there are other layers of biological complexity, such as gene expression and its regulation, that have to be harmonized between the hybrid’s counterparts to ensure successful functioning of a novel unified organism.

2.3 Transcriptomic aftermath of hybridization

Upon hybridization, the parental’s regulatory mechanisms of gene expression, which ultimately define the functioning of a cell, are merged within the same cellular environment. Thus, as in case of genomic incompatibilities, hybrids experience a “transcriptome shock” and have to coordinate transcriptome interactions of both parental counterparts to minimize possible incompatibilities (McClintock, 1984). Two primary questions arise in this context - what are the direct consequences and magnitude of “transcriptome shock”, and which molecular mechanisms govern coordination of clashing regulatory networks of two independently evolved species. Historically, the research of transcriptomes of hybrids was primarily focused on using these chimeric organisms as models for assessing *cis*- and *trans*- regulation of gene expression (Tirosh *et al.*, 2009; Graze *et al.*, 2012; Li and Fay, 2017; Metzger, Wittkopp and Coolon, 2017). Thus, the two aforementioned issues are still not sufficiently studied in fungi, preventing from having a clear idea about the magnitude of the gene expression shock, as well as about the mechanisms to cope with this transcriptomic shock.

Some recent studies performed in fungi and plants suggested several generic scenarios of hybridization outcome on gene expression levels (Yoo, Szadkowski and Wendel, 2013; Cox *et al.*, 2014; Wu *et al.*, 2018), which are summarized in Fig. 2.1. These scenarios include homeolog expression inheritance, homeolog expression blending and homeolog expression bias. It has to be noted that these outcomes are based on assessing gene expression profiles of hybrids and their parentals, and thus their mechanistic explanation on the level of *cis*- and *trans*- gene regulatory circuits are yet to be established. That is why a term ‘modulon’ was introduced (Cox *et al.*, 2014), which in essence represents all possible regulatory mechanisms by which gene expression can be regulated. In this context, hybridization represents a process of merging modulons that were evolved independently and which can interfere with each other in the hybrid.

Hence, the outcome of hybridization on gene expression can be assessed by the type of interference that takes place for each of homeologous genes. There are three main types of interference, which are going to result in the above mentioned scenarios of gene expression changes (Fig. 2.1).

If parental modulons are completely independent from each other (for example, due to very large genome divergence between parents), hybrid will inherit the same gene expression levels observed in parentals resulting in homeolog expression inheritance (Fig. 2.1A). On the other hand, if there is a full compatibility between modules (e.g. a modulon of one parental can freely target a gene of another parental, and vice versa), then gene expression levels of homeologs will be equalized, resulting in homeolog expression blending (Fig. 2.1B).

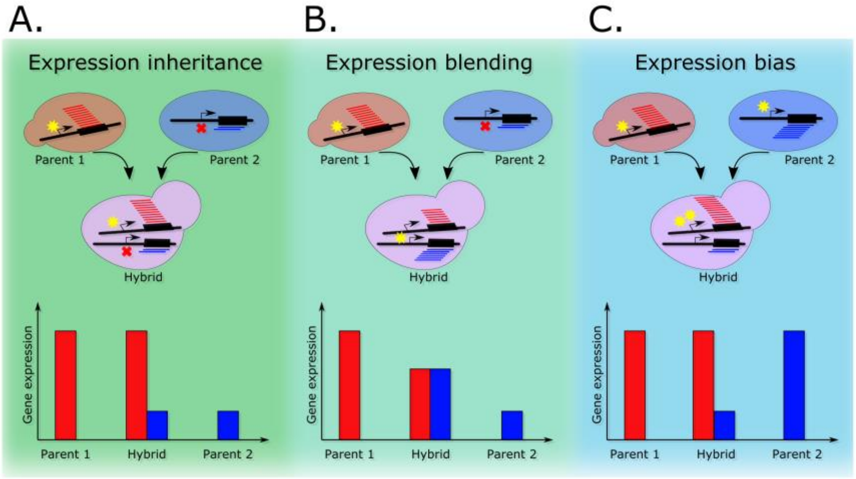


Fig. 2.1. Representation of the main generic scenarios of hybridization outcome on gene expression - homeolog expression inheritance, homeolog expression blending and homeolog expression bias. More details are given in the text. Bar plots schematically represent the outcomes of these scenarios on gene expression levels.

In the third scenario, modulon of one species is only able to target its own genes, while a module of another parent can target genes of both species. This will lead to homeolog expression bias (Fig. 2.1C). Importantly, these scenarios can coexist in the hybrid and target different portions of transcriptomes. For example, in natural fungal hybrid *Epichloë* Lp1, homeolog expression inheritance is the predominant outcome of hybridization, affecting ~56% of homeologous genes, followed by expression blending accounting for 25% of genes (Cox *et al.*, 2014). How these patterns are compared with other fungal hybrids have yet to be studied.

As stated above, these outcomes are merely based on the description of gene expression levels when comparing hybrids with their parents. Thus, we still lack sufficient mechanistic insights on how they are actually achieved in the hybrid. For example, one intriguing question is whether TF cross-talk between promoter regions of two parentals is sufficient to explain homeolog expression blending, or there are other factors, such as chromatin conformation, epigenetics, etc., that also contribute to the observed expression patterns.

Summarizing, an increasing number of studies shows that hybridization is widespread in fungi. Ancient hybridizations played a pivotal role in formation of entire fungal clades, while relatively more recent events shaped species of industrial, ecological and importantly, clinical relevance.

However, we still don’t know how the interactions between molecular systems of counterparts in a hybrid result in adaptive traits, such as virulence emergence. Some understanding of fundamental principles of these interactions on a genomic level is taking shape. However, generic rules of how hybridization influences the gene expression levels and gene regulatory circuits and how these two potentially contribute to emergence of adaptations are yet to be investigated.

3 Transcriptome sequencing approaches to elucidate host-microbe interactions in opportunistic human fungal pathogens

Hovhannisyan, Hrant, and Toni Gabaldón. "Transcriptome Sequencing Approaches to Elucidate Host–Microbe Interactions in Opportunistic Human Fungal Pathogens." *Current Topics in Microbiology and Immunology: Fungal Physiology and Immunopathogenesis* (2019), p. 193-235. DOI: 10.1007/82_2018_122

Abstract

Infections caused by opportunistic human pathogens are a cause of increasing medical concern, due to their growing incidence, the emergence of novel pathogenic species, and the lack of effective diagnostics tools. Fungal pathogens are phylogenetically diverse and their virulence mechanisms can differ widely across species. Despite extensive efforts, the molecular bases of virulence in pathogenic fungi and their interactions with the human host remain poorly understood for most species. In this context, next-generation sequencing approaches hold the promise of helping to close this knowledge gap. In particular, high-throughput transcriptome sequencing (RNA-Seq) enables monitoring the transcriptional profile of both host and microbes to elucidate their interactions and discover molecular mechanisms of virulence and host defense. Here we provide an overview of transcriptome sequencing techniques and approaches, and survey their application in studying the interplay between humans and fungal pathogens. Finally, we discuss novel RNA-Seq approaches in studying host-pathogen interactions, and their potential role in advancing the clinical diagnostics of fungal pathogens.

3.1 Introduction

Over the last two decades, the incidence of fungal infections (also known as mycoses) have increased dramatically (Pfaller and Diekema, 2007; Bitar *et al.*, 2014; Oren and Paul, 2014), particularly in hospital-associated (nosocomial) conditions (Turner and Butler, 2014; Chapman *et al.*, 2017). Fungal infections range from superficial skin lesions to life-threatening invasive infections, including fungemias. Superficial skin or mucosal infections affect around 25% of the global population (Havlickova, Czaika and Friedrich, 2008), and although they are relatively easy to manage, they collectively constitute a high burden. Invasive mycoses are life-threatening and can be associated with high mortality rates of up to 38-63%, depending on several factors such as the health status of the patient and the infecting strain (Flevari *et al.*, 2013; Klingspor *et al.*, 2015). It has been estimated that invasive fungal infections kill around 1.5 million people worldwide every year (Brown *et al.*, 2012).

The phylogenetic diversity of pathogenic fungal species is high – from estimated 2.2-3.8 million fungal species nearly 300 can cause human infections, which is likely to be an underestimation (O'Brien *et al.*, 2005; Blackwell, 2011; 'Stop neglecting fungi', 2017; Hawksworth and Lücking, 2017). Although most common fungal pathogens belong to several major clades, such as *Candida* (Pfaller and Diekema, 2007, 2010), *Aspergillus* (Dagenais and Keller, 2009) and *Cryptococcus* (May *et al.*, 2016), each of these groups can comprise numerous distinct pathogenic lineages. For example, *Candida* species, which are considered to be the most frequent and invasive opportunistic fungal pathogens (Guinea, 2014), are spread across the Saccharomycotina phylogenetic tree (Gabaldón, Naranjo-Ortíz and Marcet-Houben, 2016). *C. glabrata,* which is usually the second most abundant *Candida* pathogen, after *C. albicans* (Guinea, 2014), is phylogenetically much closer to the biotechnology workhorse *S. cerevisiae* than to other pathogenic *Candidas* (Gabaldón and Carreté, 2016). Similarly, pathogenic members of the so-called CTG clade, such as *C. albicans* (Kim and Sudbery, 2011), *C. parapsislosis* and others, have numerous non-pathogenic sister-species and clades. This phylogenetic dispersion points out that the ability to infect humans has emerged multiple times independently in genetically distinct backgrounds (Gabaldón, Naranjo-Ortíz and Marcet-Houben, 2016). Due to the high diversity and the difficulties to classify these pathogens from their physiological or morphological traits, the phylogenetic relationships between these fungi have been poorly resolved. Only the recent advent of molecular and genomic sequencing technologies has enabled accurate resolution of phylogenetic relationships, and as consequence the taxonomic nomenclature of these species is still undergoing major revisions (Brandt and Lockhart, 2012).

Novel pathogenic species are identified regularly, and the increase of incidence of previously rare species has been documented multiple times (Papon *et al.*, 2013; Short, O'Donnell and Geiser, 2014; Rhodes *et al.*, 2017). It is as yet unclear what factors drive the emergence of novel pathogens. Changes in the use of chemical products in industry or clinical care, or movement of products related to international commerce, can favor the global spread of certain species. In addition, biological factors resulting from evolutionary adaptation of microbes to novel niches can trigger the emergence of novel pathogens. One of the proposed mechanisms of emergence of novel pathogenic species in fungi is hybridization (Mixão and Gabaldón, 2018), which has been related to the formation of pathogenic lineages such as *Cryptococcus neoformans x Cryptococcus gatii* (D'Souza *et al.*, 2011), *Malassezia furfur* (Wu *et al.*, 2015), *C. metapsilosis* (Pryszcz *et al.*, 2015) and *C. orthopsilosis* (Pryszcz *et al.*, 2014; Schröder, de San Vicente, *et al.*, 2016).

Another major challenge posed by fungal pathogens is the increasing rate at which drug and multi-drug resistance (MDR) is reported (Pfaller *et al.*, 2009), which is often caused by the ability of fungi to evolve resistance phenotypes (Sanglard, 2016). The problem of resistance to one or several drugs is worsened by the limited number of available antimycotic agents, which is currently restricted to few chemical families (Kathiravan *et al.*, 2012). For example, the incidence of *C. glabrata* invasive infections has increased from 18% (in 1992-2001) to 25% (in 2001-2007), with concomitant fluconazole resistance rates increasing from 9% to 14%, respectively, in the United States (Pfaller *et al.*, 2009). *Candida auris* is another striking example of MDR pathogenic yeast, which exhibits resistance to the main classes of antifungals (Sarma and Upadhyay, 2017). Being first described in 2009 in Japan (Satoh *et al.*, 2009), *C. auris* rapidly became a notorious pathogen causing outbreaks in hospitals throughout the world (Chowdhary, Sharma and Meis, 2017). While the concern of research and medical community towards this pathogen is high, we still lack sufficient knowledge and effective approaches for controlling *C. auris* infections, highlighting the importance for public health of the emergence of antimycotic resistance.

As a consequence of high diversity, emergence of new pathogens, and increasing drug resistance, the diagnostics arsenal for the detection of the causative agent and the determination of the best treatment is limited (reviewed in Griffin and Hanson, 2014; Kozel and Wickes, 2014). Classical diagnostics methods have serious limitations. Culture-based methods using blood samples can take several days and do not provide high specificity and sensitivity, missing, for example, over 50% of the cases of documented candidiasis (Berenguer *et al.*, 1993). Moreover, some fungal species are non-culturable in conventionally used media. To overcome these issues recently novel molecular-based diagnostics tools have been developed mainly including those based on Polymerase Chain Reaction (Khot and Fredricks, 2009) or Mass Spectrometry (Chalupová *et al.*, 2014). More recently, with the rapid development of nucleic acid sequencing technologies, next-generation sequencing (NGS) might become a promising tool for microbial diagnostics (Smeekens, van de Veerdonk and Netea, 2016; Zoll *et al.*, 2016). Nevertheless, all of the aforementioned methods have their limitations from both clinical and economical perspectives (described in Griffin and Hanson, 2014; Kozel and Wickes, 2014).

Considering all these factors, it is evident that investigation of host-fungus interactions is crucial for overcoming the threats that pathogenic fungi currently impose. Firstly, knowing the specific host-evasion and virulence mechanisms used by diverse fungal pathogens may pave the way for the discovery of novel drug targets or the design of new treatment approaches. Secondly, many fungal pathogens are also commensal species that are part of the normal human microbiota. Hence, there is a need to understand what triggers may turn a commensal behavior into an invasive and virulent one. In addition, response from host cells and tissues towards different fungal pathogens may also provide important clues towards more efficient ways to avoid and control infection. Finally, understanding host-pathogen interactions may open new avenues for diagnostic approaches that are able to differentiate between commensal or infective behavior by detecting specific biomarkers. Although several studies have advanced our understanding of host-pathogen interactions for some of the more common species, the interplay between humans and fungal pathogens is, overall, still poorly understood.

NGS techniques, which allows obtaining sequence data on unprecedented scales and low costs, have represented a revolution in biological research (Goodwin, McPherson and McCombie, 2016), and the investigation of host-pathogen interaction is no exception (Hu *et al.*, 2011; Westermann, Gorski and Vogel, 2012; Westermann, Barquist and Vogel, 2017). In particular, whole transcriptome analysis by means of RNA-Seq has opened a new window to understanding gene regulation and how it changes as a result of interactions between the host and the pathogen, which potentially can shed light on the mechanisms of pathogenicity, host defense and their interplay in various conditions (Wolf *et al.*, 2018). In this review, we focus on the application of whole transcriptome sequencing in addressing host-fungus interactions during infection. We will first discuss the methodological concepts and peculiarities of RNA-Seq in the context of host-microbe interaction studies then survey past studies of human-fungus interactions based transcriptome sequencing. Finally, future perspectives in the field, including the potential of emerging technologies for the study or diagnosis of fungal infections will be discussed.

3.2 Whole transcriptome analysis methods

RNA plays a key role in the majority of cellular processes. Hence, investigation of the identity, function, and abundance of transcribed RNA molecules (i.e. transcripts) is crucial for understanding cellular behavior. Advances in the field of RNA biology were mainly driven by the development of novel technologies and methods allowing researchers to study different aspects of transcripts in an increasingly efficient way. A brief chronological overview of those techniques is discussed below, and a more in-depth comparison is provided in Table 3.1.

Initial studies of RNA molecules were performed using methods such as northern blotting (Alwine, Kemp and Stark, 1977), reverse transcriptase qPCR (Rappolee *et al.*, 1988), and expressed sequence tags (ESTs) (Adams *et al.*, 1991), which enabled to investigate individual molecules or small sets of transcripts. The first studies of transcriptomes, i.e. the whole set of RNA transcripts in a cell (or bulk of cells) at a given time point, begun in the mid 1990s, with the development of the serial analysis of gene expression (SAGE) method (Velculescu *et al.*, 1995), based on Sanger sequencing technology (Sanger, Nicklen and Coulson, 1992). SAGE and its derivatives (e.g LongSAGE, RL-SAGE, SuperSAGE) in the beginning of 2000s were largely replaced by fluorescent hybridization-based RNA microarray technologies (Schena *et al.*, 1995), which proved to be more cost-effective, as compared to previous methods. Finally, in the late 2000s, the advent of NGS superseded microarrays by RNA-Seq which provided unprecedented levels of resolution in a high throughput, unbiased, and relatively cheap manner (Bainbridge *et al.*, 2006; Wang, Gerstein and Snyder, 2009). In addition to considerations of throughput and cost, RNA-Seq presented the advantage over microarrays in that it did not require the design of probes, and could explore the entire transcriptome in an unbiased manner, enabling the discovery of novel transcripts, even in the absence of a reference genome. In the last decade, RNA-Seq has been further developed, incorporating longer sequencing reads as well as increasing its versatility by being coupled to other approaches such as, for instance, targeted enrichment (Amorim-Vaz *et al.*, 2015) or structure-specific digestion (Wan *et al.*, 2013; Saus *et al.*, 2018). Today, RNA-Seq is the major method used in transcriptomics studies.

Table 3.1. Comparison of different transcriptomics technologies

	EST/SAGE	**Microarray**	**RNA-Seq**
First description	Sutcliffe *et al.*, 1982; Velculescu *et al.*, 1995	Schena *et al.*, 1995; Lockhart *et al.*, 1996	Bainbridge *et al.*, 2006; Wang, Gerstein and Snyder, 2009

Major technology	Sanger-based sequencing of short random fragments (~ 500bp)	Hybridization to complementary probes	Next-generation sequencing
Period in active usage (Lowe *et al.*, 2017)	1994-2004	2000-2014	2009-Present
Popularity (Lowe *et al.*, 2017)	Almost obsolete	Decreasing	Increasing
Throughput	Low/Middle	High	High
Major applications	Gene expression profiling	Gene expression profiling	Gene expression profiling, discovery of novel transcripts/isoforms of any length, gene fusion, variant detection
Discovery and quantification of novel transcripts	Yes	No	Yes
Dynamic range	Up-to $3x10^5$ (Morrissy *et al.*, 2009)	10^3-10^4 (Black *et al.*, 2014)	>10^5 (Black *et al.*, 2014)
Reproducibility (R^2)	0.96 (Dinel *et al.*, 2005)	0.99 (Chen *et al.*, 2007; SEQC/MAQC-III Consortium, 2014)	0.99 (Marioni *et al.*, 2008; SEQC/MAQC-III Consortium, 2014)
Complexity of data analysis	Middle	Middle, includes image processing and differential	High (mainly command line), includes multiple steps depending on specific goal

		expression analysis	
Complexity of laboratory procedures	High	Low	High

As many other NGS-based techniques, RNA-Seq comprises two major steps: the first includes the study design and sequencing of the samples, whereas the second includes all downstream bioinformatics analysis. With current technologies the particularities of each of the two stages can vary significantly depending on the main goal of the study. Hence, there is no universal procedure for addressing all possible biological questions that can be approached with a transcriptomics approach (Conesa *et al.,* 2016). Nevertheless some generalities can be drawn. In the following sections we will focus on the main principles of each of the two steps, underscoring, when applicable, the peculiarities that are most relevant for host-fungus interaction studies. We will first discuss steps usually performed for the dominating sequencing-by-synthesis NGS technology, implemented by Illumina, while other emerging approaches, such as nanopore and PacBio sequencing, will be discussed later in our review.

3.2.1 Study design

Study design refers to the initial set up of the project, which is a crucial prerequisite for any RNA-Seq study. The project has to be planned carefully according to its main goals and taking into account the peculiarities of the addressed biological problem. Formally, the study design can be divided into experimental design and sequencing design.

3.2.1.1 Experimental design

Experimental design refers to the overall type of the study (“control-vs-treatment”, time-course, observational study, and different combinations of these types) and how it is planned and performed from both logical and technical perspectives. A poorly planned experimental design can possibly result in spurious and/or misleading results. For instance in a “control-vs-treatment” *in vivo* study of the influence of antimycotic drug on the gene expression of the fungus, “control” and “treatment” cases need to be selected. If “control” samples were obtained only from young people whereas the “treatment” sampled came from significantly older people, then the age of the donors could be a potential confounding factor, preventing from distinguishing whether the observed gene expression changes in fungus were due to the antimycotic agent or the age of the host. To avoid such kind of confounding effects, study donors for both control and treatment group have to be as similar as possible from various perspectives, controlling factors such as age, sex, diet, or the presence of concomitant diseases. Another example illustrating poorly planned experimental design could be a time-series study of a fungal pathogen interacting with host cells *in-vitro*. Without performing time-matched controls for the fungus it might not be possible to differentiate between the effect of the host-microbe interactions from the potential effect of time or growth of the fungus in the given medium. For instance, some nutrients in the medium may be exhausted, triggering physiological changes in the fungal cells, which may be wrongly interpreted as a consequence of interaction with the host. To overcome this limitation, ideally host cells-free controls have to be made at the corresponding time points of the experiment (Fig. 3.1).Thus, as illustrated by the examples above, a well-planned experimental design is a crucial first step for a meaningful RNA-Seq-based study. Critically assessing previous studies investigating similar questions (perhaps on other pathogens or hosts) and extensive discussions between all project partners (e.g. clinical doctors treating the patients, personnel collecting the samples, personnel responsible for the statistical comparisons, etc) can help to achieve a good planning and avoid potential design flaws.

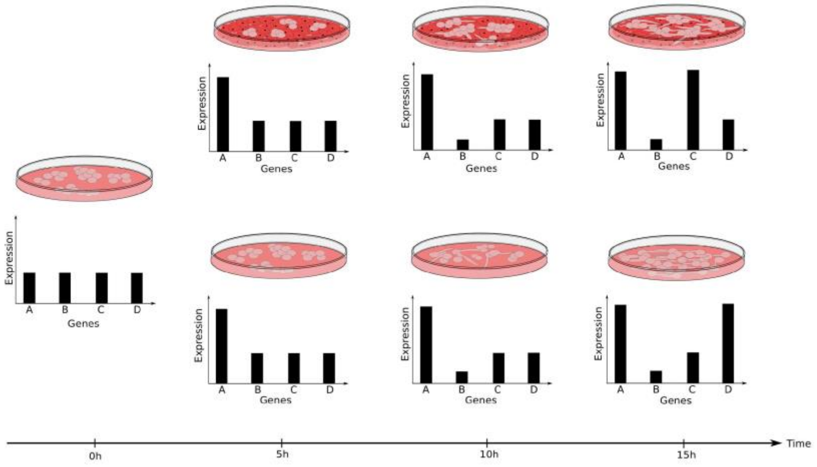


Fig. 3.1. A graphical representation of a time-series experimental design of interaction between human cell line and a pathogenic fungus. Left-most sample

represents a zero time point control, from left to right samples – 5h, 10h and 25h time-points are the host-fungus interacting samples (above) and time-matched host cells-free fungal controls (below). Bar-plots represent the expression levels of four fungal genes A, B, C and D. The overall scheme illustrates the importance of time-matched fungal controls in the experimental design: in case of their absence gene A at time point 5h, 10h, 15h and gene B at time points 10 and 15 and gene could be spuriously interpreted as up- and down-regulated, respectively, as a consequence of host-fungus interaction. Controls samples allow to distinguish the effect of media and time from the effect of host-pathogen interaction, showing that only gene C is specifically up-regulated at 15h as a consequence of interaction.

3.2.1.2 Sequencing design

The design of the sequencing approach itself refers to the main factors to consider for sample collection, storage, preparation and sequencing *per se*. While sample collection and storage methods heavily depend on a particular project, a general recommendation is to perform these two steps in the same way for all samples within a project to avoid possible confounding effects. For further sequencing the important aspects to consider are RNA extraction protocol, library preparation protocol, read type and length, number of replicates, sequencing depth, randomization of samples and sequencing runs. The combination of the aforementioned parameters entirely depends in the specific goal of the project, and some general recommendations are given in Table 3.2 and are discussed in more details in the text.

Table 3.2. General recommendations of sequencing design based on addressed biological question

Type of analysis	**Minimal recommended sequencing depth (in mln of aligned reads)**	**Read length**	**Read layout**	**Replicates**	**Popular software**	**Additional information**
Differential gene expression (DGE)	>10 mln (Liu, Zhou and White, 2014)	50bp can be sufficient	Single-end can be sufficient	>3	Mapping – STAR, HISAT2, TopHat2; Pseudomapping –	Provided in the text

analysis					kallisto, Salmon; Differential expression analysis – DESeq2, edgeR, limma-voom	
Differential isoform usage and alternative splicing (AS) analysis	>100 mln reads (Liu *et al.*, 2013)	>50 bp, longer reads are recommended	Paired-end is recommended	>3	Mapping– STAR, HISAT2, TopHat2; Alignment-free – kallisto, Salmon; Differential expression analysis – DEXseq, MISO, Cuffdiff	Read length plays critical role in finding alternative isoforms
De novo assembly and novel transcript discovery	20-60mln ** (Francis *et al.*, 2013)	>50 bp, longer reads are recommended	Paired-end is recommended	1***	Trinity, Oases, SOAPdenovo-Trans, StringTie, Cufflinks	Sequencingparameters for performing *de novo* assembly depend heavily on the complexity of the of the transcriptome (i.e. levels of AS)
Allele-specific expression (ASE)	>55 mln (not well established) (Castel	>50 bp, longer reads are recommended	Paired-end is recommended	>3	Mapping– STAR, HISAT2, TopHat2, ASE – MBASED	In general allele-specific expression analysis based on

	et al., 2015)				, Phaser, ASERead Counter (GATK)	RNA-Seq data is a relatively new area, and most of the software require variant calling file to perform ASE and only perform SNP level analysis
Small RNA-Seq (identification of miRNA, tRNA, snoRNA, etc)	>1-8 mln (Metpally *et al.*, 2013; Campbell *et al.*, 2015)	50bp can be sufficient	Single-end can be sufficient	>3	Mapping – Bowtie, BWA; Identification – miRDeep, miRDeep2, miRanalyzer	As a rule, discovery of small RNAs requires more sequencing depth than differential expression analysis

*Numbers are given for human transcriptome. For differential expression analysis, given numbers of reads are applicable for highly and moderately expressed transcripts
** The data is for non-model organisms
*** In general case replications are not strictly necessary as in DGE analysis, but are recommended for reproducibility of the results
mln = million

For fungi, several commercial kits are available for high quality and high yield RNA extraction. An important factor to consider on this step is whether rRNA depletion or poly-(A) selection is required, since usually transcriptome studies are focused on mRNAs, while ribosomal RNA can constitute the vast majority of RNA in a cell (i.e. up to 60% in an exponentially growing *Saccharomyces cerevisiae* cells (Warner, 1999)), and its removal will provide higher resolution for the mRNAs (O’Neil, Glowatz and Schlumpberger, 2013; Zhao *et al.*, 2014). Moreover, different strategies to enrich specific mRNA molecules have been recently developed, which will be discussed in more detail in the following section of our review.

The specific sequencing library preparation protocol is yet another factor to consider, especially with regards to its ability to generate strand-specific data. Early library preparation protocols were not capable of preserving information about the strand of DNA from which a transcript originated, thus biasing, for example, gene expression analysis by anti-sense transcription (Zhao *et al.*, 2015). Today several so-called strand-specific (or stranded) protocols are available, such as dUTP (Parkhomchuk *et al.*, 2009; Borodina, Adjaye and Sultan, 2011), RNA-ligation (Lister *et al.*, 2008), SMART (Zhu *et al.*, 2001) for retaining strand information of transcripts. A comprehensive benchmark of these protocols is given in (Levin *et al.*, 2010).

The length of the reads (short stretches of cDNA that are actually sequenced) is one of the major parameters of the sequencing design. The read length of Illumina-based sequencing varies from 25 to 300 bases depending on the model of sequencing machine. As an additional option to increase the capabilities of obtained data, one can perform paired-end (PE) sequencing instead of single-end (SE). In the former case the cDNA fragment is sequenced from both ends, hence doubling the amount of the information obtained from it. As a general rule, longer reads coupled with paired-end long-insert size sequencing provide with higher mapping rates to reference genomes, more accurate transcript discovery, the ability to detect larger indels, among other advantages. However, PE sequencing and longer reads come with higher price, which may compromise the number of replicates if budget is limited. Moreover some particular goals could be sensitively achieved even with short SE reads (Chhangawala *et al.*, 2015). For example, considering that in *Candida albicans* introns are not abundant in the genome and their length is usually short (Mitrovich *et al.*, 2007; Chhangawala *et al.*, 2015) (Mitrovich et al. 2007), long PE reads are not critical for most of the typical downstream analysis.

In the context of typical RNA-Seq applications, such as differential gene expression (DGE), replicates and sequencing depth are two crucial and interconnected parameters. When choosing the number of biological replicates one has to consider the intrinsic biological variability of the studied system, the technical variability of the experimental procedures, and the desired statistical power of the experiment. As a general rule, the number of biological replicates included in the study should be at least 3, while the recommended number ranges from 6 to 12 biological replicates (based on *S. cerevisiae* data) depending on the specific goals of the study (Schurch *et al.*, 2016). Several approaches and calculators have been implemented to perform RNA-Seq power analysis to help deciding the number of replicates in an RNA-Seq design (Hart *et al.*, 2013; Guo *et al.*, 2014; Yu, Fernandez and Brock, 2017).

Sequencing depth (or library size) denotes the total number of reads for each sample to be sequenced. Higher sequencing depth allows more precise transcript detection and expression quantification, but also might suffer from transcriptional noise and false-positive calls of DGE (Tarazona *et al.*, 2011). Thus, once again the optimal sequencing depth depends on the addressed question and the system under study.

As mentioned above, replication and library size are interconnected parameters and for a particular sequencing design one can wonder whether for a fixed budget it is more advisable to add more replicates to the study or to perform deeper sequencing. Liu et al. (2014) have evaluated the impact of both factors on DGE analysis. The study revealed that, in the case of human transcriptomic data, increasing the sequencing depth over 10 million reads has diminishing incremental effects for power of detection of differentially expressed (DE) genes, whereas an increase in the number of biological replicates significantly enhances the power of detection. Authors of the study also suggest a metric of cost-effectiveness of RNA-Seq design as a trade-off between number of replicates and sequencing depth. The last but no least step in experimental design in the case of projects considering a high number of samples is to randomize the distribution of samples on different sequencing lanes or runs. This step is meant to avoid possible confounding factors such as different instrumental biases (lane effects, PCR duplicates, etc) as well as difficult to control human factors.

3.2.2 Bioinformatics data analysis

As any high-throughput sequencing technology, RNA-Seq generates massive amounts of data (i.e. a typical DGE RNA-Seq analysis of yeast with 2 conditions and 3 replicates yields at least 100-150 million reads) which have to be thoroughly analyzed using bioinformatics approaches. Considering that RNA-Seq has numerous applications, the complete pipeline for bioinformatics analysis varies depending on the specific goals of the project. With regards to host-pathogen interaction studies, one of the most frequent RNA-Seq application is the analysis of differential gene/transcript expression (Westermann, Barquist and Vogel, 2017). Here we will briefly describe the main steps of this bioinformatics approach (Fig. 3.2).

The general initial step for any NGS-based data analysis is quality control (QC) of the raw data produced by the sequencing machine. As a general rule, raw reads are stored in the standard *.fastq* format (Cock *et al.*, 2010), which provides associated per base quality scores. For basic QC several software solutions can be used, such as FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), HTQC

(Yang *et al.*, 2013), or NGS QC (Patel and Jain, 2012). The main parameters to assess include, among others, per base sequence quality, GC content, or the presence of overrepresented sequences and/or those corresponding to the library adapters. When some quality parameters are not satisfactory, one can perform several actions including read trimming to cut-out low quality reads or bases, or removing adapter sequences coming from the library preparation step. Popular software to perform read trimming are Trimmomatic (Bolger, Lohse and Usadel, 2014), Skewer (Jiang *et al.*, 2014), among others. However, in the case of DGE analysis trimming has to be performed gently since the harsh trimming can affect the results of read mapping and differential expression calls (Williams *et al.*, 2016). For this reason, if the quality of the data is still unsatisfactory after trimming, it is advisable to resequence the sample or the library.

After ensuring the high quality of the sequencing data, downstream analysis depends on the presence or absence of a reference genome or transcriptome of the studied organism. In the case of human fungal pathogens, most common species like those belonging to *Candida*, *Aspergillus*, or *Cryptococcus* groups, have available reference genomes in both specialized and generic public databases such as *Candida* Genome Database (Binkley *et al.*, 2014), *Aspergillus* Genome Database (Cerqueira *et al.*, 2014), FungiDB (Stajich *et al.*, 2012), RefSeq (Pruitt, Tatusova and Maglott, 2007), etc. Alternatively, in the case of absence of reference, the transcriptome can be reconstructed *de novo*, using, for example, Trinity package (Haas *et al.*, 2013) or SOAPdenovo-Trans (Xie *et al.*, 2014).

When a reference genome or transcriptome is available, downstream bioinformatics analysis implies mapping of reads to this reference. This computationally demanding task can be achieved by splice-aware read mappers (despite the fact that splicing is not as common in easts as in more complex eukaryotes). Numerous RNA-Seq mappers exist today, and choosing one might not be a trivial task. In fact, many researchers have addressed this question by performing benchmarks comparing different mappers, sometimes reaching contradicting conclusions (Dobin and Gingeras, 2013; Kim *et al.*, 2013; Otto, Stadler and Hoffmann, 2014; Baruzzo *et al.*, 2017). Nevertheless, the most popular mappers include STAR (Dobin *et al.*, 2013), TopHat2 (Kim *et al.*, 2013), HISAT2 (Kim, Langmead and Salzberg, 2015). After reads are mapped to the reference, quality control of the overall mapping is required.

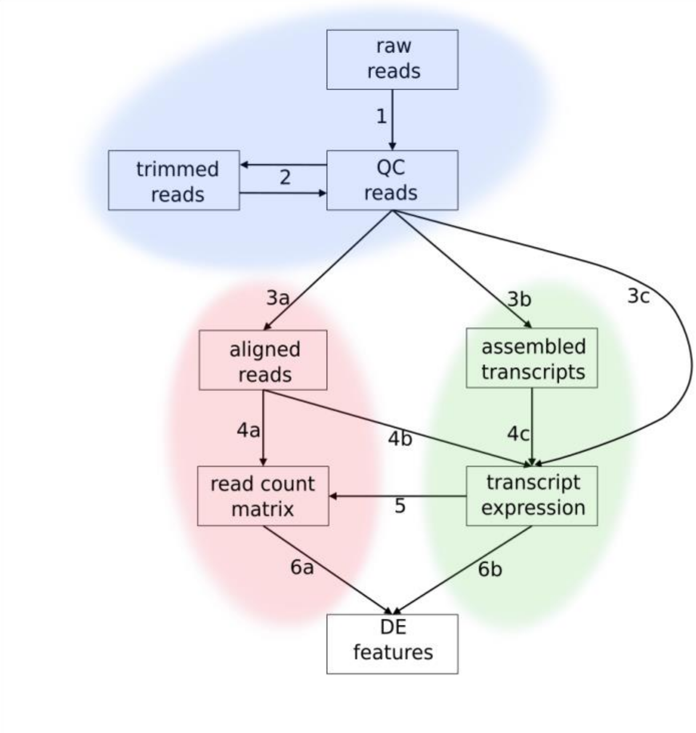


Fig. 3.2. General representation of RNA-Seq differential expression analysis pipeline. The numbers correspond to different steps of the analysis: 1 – quality control of raw data; 2 – read trimming (if necessary), 3a – read mapping to a reference genome, 3b – *de novo* assembly of transcripts 3c – pseudomapping strategy (requires reference transcriptome), 4a – read summarization, 4b – reference-guided transcriptome assembly (used for transcript identification), 4c – transcript quantification, 5 – transcript level quantifications can be converted to gene level count, which improves gene-level inferences, 6a – differential gene/transcript/exon/feature analysis based on read counts, 6b – differential gene/transcript/exon/feature analysis based on relative expression values. Shades indicate different major strategies of the analysis: blue – general step of raw data quality control and read trimming; pink – down-stream analysis when the reference genome is available, mainly includes gene level inferences, green – when the reference genome is not available and includes transcript level inferences.

Most software tools provide basic mapping statistics, which include, among others, overall mapping rate, unique mapping rate, rate of multimappers (reads that map equally well to different locations in the reference), number of identified splices, etc. Unique mapping rate is one of the crucial parameters, and usually with high quality raw data and a good enough reference genome the optimal values range from 85% to 95% of uniquely mapped reads. If the value is significantly lower, it might indicate low quality of reads, poor quality of reference genome assembly or the nature of reference genome itself – for instance, genomes with large amount of repeats might result in increased numbers of multimapped reads. The next step after read mapping is read summarization, which involves the calculation of the number of reads that overlap particular genes, which is proportional to the expression levels. This procedure relies on gene annotations (gff or gtf files) and popular software to accomplish this task are htseq-count (Anders, Pyl and Huber, 2015) and featureCounts (Liao, Smyth and Shi, 2014), with the latter being more flexible in dealing with multimapped reads.

The two previous steps described gene-level analysis to obtain the information about expression values based on genome alignments. However, today several so-called pseudo-alignment algorithms are available that allow the assessment of expression levels of individual transcripts, rather than of genes. Examples of such algorithmic implementations include Salmon (Patro *et al.*, 2017) and kallisto (Bray *et al.*, 2016). For organisms with high rates of alternatively spliced transcripts it has been recently demonstrated that transcript-level estimates could improve gene-level inferences (Soneson, Love and Robinson, 2015). The final step of DGE bioinformatics analysis is the assessment of DGE between groups of samples. Many studies have been performed to evaluate the most effective models and corresponding software for assessing differential expression (Bullard *et al.*, 2010; Rapaport *et al.*, 2013; Soneson and Delorenzi, 2013; Seyednasrollah, Laiho and Elo, 2015; Schurch *et al.*, 2016). Readers are referred to (Conesa *et al.*, 2016) for more details on this matter. When sufficient number of biological replicates (3 or more) are available, software tools such as DESeq2 (Love, Huber and Anders, 2014), edgeR (Robinson, McCarthy and Smyth, 2010) and *limma (Ritchie et al., 2015)* generally perform well in most of circumstances. To choose the genes that are differentially expressed (DE), one has to set a cut-off for both fold-change of gene expression between conditions and p-value of statistical significance of this change, and usually these thresholds are arbitrary and depend on the desired level of stringency.

Since overall RNA-Seq analysis consists of many wet-lab and data analysis steps, it is advisable to confirm a subset of the obtained results with an alternative approach. For instance one can perform quantitative PCR on a subset of DE genes to ensure the reliability of the obtained results. Once the DE genes have been identified, researchers can proceed with further in-depth analysis, which, among others, usually includes gene ontology (GO) (The Gene Ontology Consortium, 2017) or gene set enrichment analysis (GSEA) (Subramanian *et al.*, 2005), pathway analysis (Emmert-Streib and

Glazko, 2011) and gene co-expression and network analysis (Schulze *et al.*, 2016). GO and GSEA are two different approaches addressing a similar question – whether DE set or subset of genes is enriched in a specific biological function, process or cellular location. For instance, when differential expression analysis reveals hundreds of DE genes, these methods help to get an overall insight which biological functions are altered in a given condition. Similarly, pathway analysis allows to identify specific molecular pathways which are dysregulated in the studied system. As in GO and GSEA, pathway analysis is performed based on statistical test of enrichment between sets of gene lists.

Gene co-expression and network analysis on the other hand allows to quantitatively assess genes which are changing the expression levels systematically in a similar manner, revealing gene-gene interactions. Especially this kind of analysis is relevant when a time-series dual RNAseq data is available, which enables to detect the interacting co-expressed genes of the host and the pathogen (Schulze *et al.*, 2015). For a more detailed discussion about the best practices of RNA-Seq related analysis, including both study design and bioinformatics data analysis, readers are referred to the recent review by Conesa et al. 2016.

3.2.3 Dual RNA sequencing

Dual RNA sequencing (dual RNA-Seq) is a relatively new methodology (Westermann, Gorski and Vogel, 2012) of simultaneous sequencing of RNA that originates from two (or more) organisms. Originally dual RNA-Seq was developed in the context of host-pathogen interaction studies, allowing to profile the gene expression of both counterparts at the same time, but in principle can be used to study the interactions between any cohabiting organisms. The main idea behind this method is to sequence the mixture of RNA, that contains transcripts from two or more organisms. The mixture of RNA could be obtained by direct extraction of RNA from both species (e.g. when studying interaction between co-cultured bacterial species) or it can be extracted separately for each species and then mixed into one sample. The latter approach is more suitable for host-microbe interaction studies, since RNA extraction protocols for fungi/bacteria include a cell-wall disruption step, which can degrade the RNA content of the host cells. After sequencing such a mixed sample, the reads from both species are separated bioinformatically by mapping the mixture of reads to both reference genomes simultaneously. When the reads are successfully separated, the data analysis is largely the same as in the case of standard RNA-Seq. Hence, the above mentioned recommendations of experimental and sequencing design for common RNA-Seq are also applicable in dual RNA-Seq.

Despite the fact dual RNA-Seq methods are in its infancy, they have already been proven to be an efficient tool for dissecting the interplay between hosts and pathogens (Bruno *et al.*, 2015; Aprianto *et al.*, 2016; Dutton *et al.*, 2016; Nuss *et al.*, 2017; Thänert *et al.*, 2017). Nevertheless, this approach has some technical limitations that need to be overcome. Firstly, in *in-vivo* studies, particularly those involving fungi, the amount of microbial cells and its corresponding RNA is extremely low, as compared to the host side. The RNA-Seq of the sample, heavily shifted towards the host, yields a negligible amount of fungal reads, precluding detailed analysis of fungal transcriptome. To date, the most efficient way to overcome this problem is by using targeted enrichment of fungal transcripts (Amorim-Vaz *et al.*, 2015), which we discuss later. Second, since the dual RNA-Seq method is new, specific software and data-analysis pipelines still do not exist. The major bioinformatics problem that can arise in dual RNA-Seq experiments is cross-mapping of reads to the wrong reference genome, since the mixture of reads is mapped to both references simultaneously, biasing downstream analysis. Thus, specific data analysis pipelines should be implemented in order to remove that kind of reads. Despite these difficulties, dual RNA-Seq holds a great potential in resolving interactions between species on a transcriptome-wide manner.

3.3 RNA-Seq based studies to understand human-fungus interactions in *Candida*, *Aspergillus* and *Cryptococcus* clades

RNA-Seq has emerged as a versatile tool for studying host-pathogen interactions at the transcriptomic level. The majority of transcriptomic studies for elucidating pathogenic mechanisms in fungi so far has been performed *in-vitro* by exposing the pathogen to different experimental conditions that try to mimick stress factors encountered in the host. These include, among many others, low pH, oxidative stress, or different temperatures (Lin *et al.*, 2013; Cottier *et al.*, 2015; Brown *et al.*, 2016; Yang *et al.*, 2016; Cheon *et al.*, 2017). However, a limited number of transcriptomic studies have been performed *in vivo*, readily characterizing transcriptome responses of the pathogen, host or both during their direct contact as it takes place during a real infection. Although this approach faces numerous challenges, it is still crucial for disentangling genuine human-fungal interactions. Here, we provide an overview of significant insights gained from transcriptomic studies. For simplicity, we will focus on the three major clades of fungal pathogens, namely *Candida*, *Cryptococcus*, and *Aspergillus*, as *r*esearch on other fungal pathogens generally lag behind. A schematic summary of surveyed studies is given in Figure 3.3.

3.3.1 *Candida*

The most well-studied opportunistic pathogen from *Candida* species is *Candida albicans*, and its virulence mechanisms and host-fungus interactions have been extensively reviewed in Wilson and Hube (Wilson and Hube 2014). Briefly, primary pathogenic mechanisms of *C. albicans* explored to date include hyphae formation (Sudbery, Gow and Berman, 2004) alongside with the expression of virulence factors, such as candidalysin (Moyes *et al.*, 2016), adhesins (e.g. *HWP1*, *HGT2*) (Nobile *et al.*, 2006; Martin *et al.*, 2013), invasins (e.g. *ALS3*) (Liu and Filler, 2011) and secreted proteases (e.g. *SAP4-6*) (Naglik, Challacombe and Hube, 2003). Hyphae formation and the expression of virulence factors promote initial adherence to the host tissue followed by invasion (either induced endocytosis or active penetrations) and damage. In turn, host defense against infecting *Candida* is mainly presented by the action of macrophages and neutrophiles (reviewed in Moyes et al., 2014; Wilson and Hube 2014). After phagocytosing the fungal cell, neutrophils expose a variety of factors to block hyphae formation and eventually kill the fungus, including nutrient starvation, production of antimicrobial peptides and enzymes (e.g. defensins, lactoferrin, ellastase), oxidative burst, formation of neutrophil extracellular traps (NETs), etc.

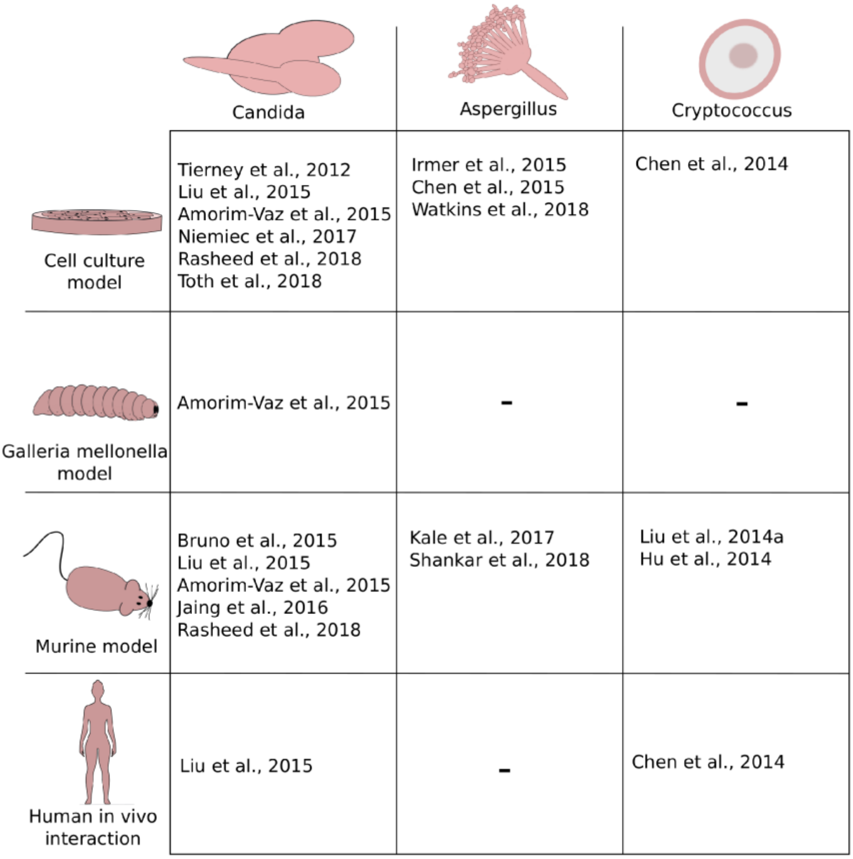
As mentioned above, quantities of yeast cells in an infected patient sample are generally very small, which poses many challenges for the analyses (Rosenbach *et al.*, 2010; Bruno *et al.*, 2015). As a consequence many previous studies have been performed using animal or tissue culture models, where higher loads of the pathogen can be present and larger quantities of tissue are available. One of the first studies using RNA-Seq to decipher host-pathogen interactions of *C. albicans* was carried out by Tierney et al (Tierney *et al.*, 2012). In this study, the authors performed an *in-vitro* time-course model experiment of interaction between *C. albicans* and *M. musculus* bone marrow-derived dendritic cells (BMDCs) with further RNA sequencing and network analysis to identify and predict interspecific interactions.

With the aforementioned techniques, the authors predicted and subsequently experimentally verified a mechanism by which *C. albicans* escapes host immune response mediated by a reorganization of its cell wall, which in turn is triggered by the release of complement-activating and opsonin protein Ptx3 from dendritic cells

In a more recent study Bruno et al (Bruno *et al.*, 2015) used a murine model of vulvovaginal candidiasis (VVC) coupled with RNA-Seq to study the transcriptome and its alterations in mice and *C. albicans*. This study demonstrated that expression of the NLRP3 inflammasome, which triggers caspases and the maturation of proinflammatory cytokine interleukin-1beta – the hallmark of VVC immunopathogenesis, was elevated in infected mice. Moreover, *Nlrp3−/−* infected mice showed significantly lowered levels of polymorphonuclear leukocytes (PMNs), alarmins, and inflammatory cytokines. These findings suggested an important role of NLRP3 inflammasome in response to *C. albicans* in VVC. On the other hand, the authors have also attempted to analyze *C. albicans* response to host, however the in-depth analysis of *C. albicans* transcripotme was precluded by a very low amount of fungal reads obtained from infected vaginal samples (on average ~80 thousand reads of *C. albicans* compared to ~103 million mouse-derived reads). Nevertheless, the analysis of highly expressed *C. albicans* genes revealed a robust expression of hypha-associated *SAP 4*, *5* and *6*, while mutants of these genes were inducing significantly lower inflammatory response.

Another comprehensive RNA-Seq-based study of *C. albicans* and host interaction was done by Liu et al (Liu *et al.*, 2015). Here the authors analyzed host-pathogen interactions in both *in-vitro* and *in-vivo* conditions. In the former case two human cell cultures were used, namely human endothelial and oral epithelial cell cultures, while *in-vivo* investigation was carried out in both murine model and using real clinical samples from patients. The time-series analysis using not only infected samples but also controls on each corresponding time-point (non-infected human samples and *C. albicans* cells in the growth media) allowed the researchers to reveal that, surprisingly, *C. albicans* showed a minimal transcriptional response to host cells, which was indicated by a low number of DE genes compared to the control samples. This result points out that most of the *C. albicans* genes involved in host interaction with studied cell types are also similarly expressed in the growth medium (M199 and DMEM media). Nevertheless, this fact does not lessen the importance of these genes in host-pathogen interaction, but rather once again demonstrates the relevance of careful study design. Thereafter focusing mainly on the host side the authors demonstrated a distinct transcriptional response of different cell lines to *C. albicans*.

To identify molecular pathways governing host response, the authors used network analysis and identified numerous previously reported up-regulated pathways like MAPK1/3, TLR7, EGF, and novel pathways such as PDGF and NEDD9. Further in-depth wet-lab analysis has shown that the last two pathways play a crucial role in endocytosis of *C. albicans* cells in a cadherin-independent manner in cell cultures.



	Candida	Aspergillus	Cryptococcus
Cell culture model	Tierney et al., 2012 Liu et al., 2015 Amorim-Vaz et al., 2015 Niemiec et al., 2017 Rasheed et al., 2018 Toth et al., 2018	Irmer et al., 2015 Chen et al., 2015 Watkins et al., 2018	Chen et al., 2014
Galleria mellonella model	Amorim-Vaz et al., 2015	-	-
Murine model	Bruno et al., 2015 Liu et al., 2015 Amorim-Vaz et al., 2015 Jaing et al., 2016 Rasheed et al., 2018	Kale et al., 2017 Shankar et al., 2018	Liu et al., 2014a Hu et al., 2014
Human in vivo interaction	Liu et al., 2015	-	Chen et al., 2014

Fig 3.3. Host-pathogen interactions transcriptomic studies of *Candida*, *Aspergillus* and *Cryptococuss* species in different experimental models.

Moreover, it was demonstrated that both pathways are also implicated in pathogen interaction in the disseminated mouse infection model, while in case of *in-vivo* human infection NEDD9 was intact. Overall, this study is an excellent example where a combination of carefully planned RNA-Seq design and thorough bioinformatical and follow-up wet-lab analysis can successfully reveal novel mechanisms of host-fungal interaction.

A recent study by Niemiec and colleagues (Niemiec *et al.*, 2017) has evaluated the interaction between *C. albicans* and human neutrophils by means of RNA-Seq. The authors assessed the transcriptome of neutrophils exposed to the fungus in either yeast and hyphal morphotypes, as well as the transcriptomic response of those morphotypes to intact neutrophils and NETs. The analysis revealed that the core response of neutrophils is largely similar for the two *C. albicans* morphotypes, with only 11% of DE genes being specific to the interaction with the hyphal morphotype. The core response to fungi included inflammasome induction and release of numerous cytokines, which shows that despite their short life-span neutrophils are also orchestrating complex immune response to *C. albicans*. On the other hand, *C. albicans* response was also mainly morphotype-independent, while the reaction to either intact neutrophils or NETs was markedly distinct. Overall, fungal response was primarily dominated by metabolic genes, controlled by the regulators of transcription as Tup1p, Cap1p, Hap43p, with the latter being the major regulator in *C. albicans* of evasion from neutrophils.

As highlighted previously, one of the major limitations of investigating mutual host-fungus interaction (especially in *in vivo* studies) is a very low proportion of fungal cells as compared to host cells (Rosenbach et al. 2010; Bruno et al. 2015). This problem refers not only to RNA-Seq, but to any other high-throughput NGS technique. Subsequent analysis of such kind of a “host-biased” sample generally does not yield enough fungal data for a comprehensive description of the fungal transcriptome. Previous attempts to solve this issue had serious limitations – some of them were altering true gene expression levels (Andes *et al.*, 2005; Thewes *et al.*, 2007), while the others did not provide transcriptome-wide resolution (Geiss *et al.*, 2008). To overcome this issue Amorim-Vaz et al. (Amorim-vaz et al. 2015) used, for the first time, RNA-Seq coupled with SureSelect targeted RNA enrichment technology. This technology is based on the use of biotinilated oligonucleotide baits directed to target RNA molecules of interest, which are then enriched by probe hybridization and subsequent pool down. Importantly, this enrichment procedure has been shown to not interfere or change the transcriptional profile of the sample. The authors used two animal models of *C. albicans* infection – murine model of kidney infection and a *Galleria mellonella* larvae model – and investigated host-pathogen interplay at early and late stages of infection. Consistent with previous studies, the RNA-Seq analysis of the bulk sample showed that barely 0.1-1% of the reads belonged to the pathogen. However, after applying the enrichment approach the number of reads aligned to *C. albicans* dramatically increased up to 1,670-fold, while biasing the expression levels only for 3% of genes. DGE analysis of *C. albicans* genes showed consistent results with previously published studies including up-regulation of genes involved in cell host adhesion, hypha formation, and iron acquisition. Moreover, the analyses revealed new, previously uncharacterized targets in both *C. albicans* and hosts for further exploration. Overall, this study demonstrated that targeted enrichment can be successfully applied for *in-vivo* host-pathogen studies to describe both counterparts in a transcriptome-wide manner.

3.3.1.1 Non-albicans *Candida* species

Other *Candida* species are less frequently reported than *C. albicans* in infection cases, but nevertheless they collectively account for ~50% of the cases. After *C. albicans*, the most widespread species in infections are *C. glabrata*, *C. parapsilosis* and *C. tropicalis,* generally in this order (Guinea 2014). Despite the importance of these species in fungal infection epidemiology, their virulence mechanisms and host interactions are significantly less studied than those of *C. albicans*. So far, only a handful of studies have been performed for transcriptional profiling of these species in the context of host-pathogen interactions, and none of them was deeply focused on both counterparts.

Candida glabrata is the second most widespread *Candida* species that causes human infections. Phylogenetically, this yeast is much closer to *S. cerevisiae* than to *C. albicans,* it does not form true hyphae and has high intrinsic resistance to azole class of antifungal drugs (Gabaldón and Carreté 2016). Rasheed et al. (Rasheed, Battu and Kaur, 2018) investigated the role of yapsins (*Cgyps*) – cell surface-associated aspartile proteases of *C. glabrata* – in the interaction with human THP-1 macrophages and in systemic murine infection. First, to clarify the role of yapsins in fungal homeostasis on gene expression level**,** the authors performed RNA-Seq of mutant strain *Cgyps1–11*Δ, which lacks all 11 yapsins, and compared it to the wild-type (WT). Downstream analysis uncovered 35 down- and 89 up-regulated genes in the mutant, with enriched GO categories of “ion transport”, “oxidation-reduction process” and “sterol import”, and “carbohydrate metabolic process”, “fungal-type cell wall organization” and “tricarboxylic acid cycle”, respectively. Using biochemical staining assays, the authors further demonstrated the altered cell wall composition of the *Cgyps1–11*Δ mutant in β-glucan, chitin and mannan content, largely caused by the deletion of *CgYps1* and *CgYps7* yapsins. While the application of RNA-Seq was restricted to the above mentioned analysis in this study, the authors additionally used microarray technology to describe human THP-1 macrophages response to *C. glabrata* WT and the forementioned mutant. Broadly, the microarray profiling showed that THP-1 cell line responds differently to WT and mutant *C. glabrata* strains: in the former case human cells DE genes involved in inflammatory response, chemotaxis, and chemokine-mediated signaling pathways, while in the latter they expressed genes involved in viral response. In addition, the authors elucidated the role of IL-1β in *C. glabrata* interaction, showing that its production is likely to be deleterious for fungal survival in macrophages, and that yapsins play a pivotal role in suppressing the production of host’s IL-1β.

To clarify the role of yapsins *in-vivo,* BALB/c mice were infected with WT and *Cgyps1–11*Δ mutant. Overall, WT *C. glabrata* colonized and disseminated in numerous mouse organs, while the mutant strain had a significantly lower survival, demonstrating that yapsins are required for colonization and dissemination of the fungus. Finally, to uncover the roles of each of the yapsins in infection and fungal survival, the authors performed murine infection models with different combinations of single, double and triple mutants of yapsin genes. Altogether, organ-specific survival effects of different yapsins were identified.

Another recent study (Whaley *et al.*, 2018) focused on *C. glabrata* addressed the susceptibility mechanisms of the fungus to fluconazole, identifying the gene which negatively regulates the resistance levels. By screening a large collection of single gene mutants, the authors found that the strain with deleted *JJJ1* gene (*GL0J07370g*), increased the minimum inhibitory concentration (MIC) to fluconazole 16 fold as compared to WT. This finding was further supported by deleting this gene in a *C. glabrata* clinical strain. Since the main mechanism of *C. glabrata* resistance to azoles is defined by over-expression of the transcription factor PDR1, which directly activates efflux-pump genes such as *CDR1*, *PDH1* and *SNQ2*, the authors demonstrated that deletion of *JJJ1* increased the resistance through Pdr1-dependent up-regulation of *CDR1*. To further investigate the effect of *JJJ1* deletion on the overall transcriptome of *C. glabrata*, the authors have performed RNA-Seq using Ion Torrent technology. The analysis identified 119 and 149 up and down-regulated genes, respectively, many of which had been previously identified by microarray analysis.

Candida parapsilosis is a member of CTG clade alongside with *C. albicans* and *C. tropicalis*. It is considered to be the third most frequent opportunistic *Candida* pathogen. As for *C. glabrata*, a restricted number of studies have been performed to clarify host-pathogen mechanisms on transcription level. To our knowledge, the only RNA-Seq-based host-pathogen interaction study with *C. parapsilosis* was performed recently by Toth et al. (Tóth *et al.*, 2018), focusing mainly on the fungal side. The authors employed a time-course *in-vitro* infection model of *C. parapsilosis* with human THP-1 monocytes with further RNA-Seq of fungal transcriptome to identify potential molecular targets for future antimycotic agents. RNA-Seq analysis revealed 19 highly up-regulated *C. parapsilosis* genes, which were selected for further investigation. By constructing deletion mutant strains of each of those genes and performing the screening of the mutants for different properties, the authors narrowed the search of virulence factors to three transcriptional regulator genes *CPAR2_100540*, *CPAR2_200390* and *CPAR2_303700*. Further in-depth analysis demonstrated that these three genes play an important role in nutrient acquisition and alternative carbon source utilization, hyphae and biofilm formation, and sensitivity to low temperatures, respectively.

As for *C. tropicalis*, two studies have been performed that used RNA-Seq to understand yeast-hyphal transition. Wu et al. (Wu *et al.*, 2016) performed RNA sequencing of three *C. tropicalis* clinical isolates in yeast and filamentous forms. Differential gene expression analysis showed up-regulation of several genes, including *SAP2, SAP3, ALS3, LIP1,* which have been previously reported to be involved in hyphal transition in *C. albicans*.

Jiang et al. (Jiang *et al.*, 2016) have studied 52 clinical isolates of *C. tropicalis* by estimating different parameters of pathogenicity, such biofilm formation, hyphal morphology and hemolytic activities. Based on the ability to form hyphae, two groups of strains (three highly and three lowly hyphae-producing strains) were further selected for performing *in-vivo* murine infection model with subsequent RNA sequencing of *C. tropicalis*. RNA-Seq analysis between two groups has shown 206 DE genes in highly hyphae-producing strains, enriched in aspartic-type endopeptidase activity, metal homeostasis and oxidative response. On the other hand, several uncharacterized DE genes were revealed, which might also have an impact on *C. tropicalis* pathogenicity.

3.3.2 *Aspergillus*

Aspergillus is a genus within the *Ascomycota* phylum that comprises over 300 species (Samson *et al.*, 2014). *Aspergillus* species have a high and diverse economic and social impact, since they massively spoil food products (Dijksterhuis, Houbraken and Samson, 2013), serve for various biotechnology productions (Pel *et al.*, 2007) and some are human pathogens (Kwon-Chung and Sugui, 2013). From the latter perspective, the most frequent human pathogen is the soil-associated fungus *Aspergillus fumigatus,* accounting for 90% of *Aspergillus*-caused infections, which as in the case of *Candida* affects mainly immunocompromised individuals (Perfect *et al.*, 2001; Paterson and Lima, 2017). *A. fumigatus* produces hydrophobic microscopic spores known as conidia, which are ubiquitous in the environment and are the main cause of infections (Latgé, 1999). After being inhaled by immunocompromised individuals, conidias can reach pulmonar alveoli and start germinating by forming hyphae and mycelia, which cause a wide range of nosologies collectively called aspergillosis, with invasive forms reaching 50-95% mortality rates (Abad *et al.*, 2010).

Numerous studies involving transcriptome profiling have been performed to study pathogenic mechanisms of *A. fumigatus*. Most of them have been carried out *in-vitro* by exposing the fungus to different environments resembling the interaction with the host and or to different stresses

(Gibbons *et al.*, 2012; Losada *et al.*, 2014; O'Keeffe *et al.*, 2014; K. Wang *et al.*, 2015). However, only few recent studies addressed the host-pathogen interactions based on RNA-Seq of more realistic *in vivo* models focusing either on host, pathogen, or both simultaneously.

Using RNA-Seq, Irmer et al. (Irmer *et al.*, 2015) investigated the response of *A. fumigatus* exposed to human blood *in vitro*, mimicking the host environment encountered by the fungus when it germinates and penetrates to blood vessels. The authors took samples at two time-points of 30 and 180 minutes after incubating with either human blood or with minimal medium, as a control. The experiment was performed in duplicates and all samples were compared to pre-cultured *A. fumigatus* mycelia. Differential gene expression analysis between pre-cultured fungus and blood-exposed samples revealed 410 up-regulated and 367 down-regulated genes after 30 minutes of exposure to blood, and 266 up-regulated and 318 down-regulated genes after 180 minutes. Those numbers of genes were obtained after subtracting the DE genes from the comparison between control samples and fungi grown on minimal media. After differential expression analysis, the authors performed comprehensive GO enrichment analysis. Briefly, 4 categories of genes were analyzed – early up-regulated genes, early up-regulated and then down-regulated genes, solely late down-regulated genes and late up-regulated genes. Functional analysis of early up-regulated genes showed the enrichment in metabolism, cell-rescue, transport, virulence and protein synthesis related genes. Genes that were up-regulated after 30 min but down-regulated after 180 min were largely similar to the ones only up-regulated at 30 min or only down-regulated after 180 min and thus their functional enrichments were also similar. After 180 min, enrichment categories remained largely the same compared to early up-regulation, but, obviously, were depressed, indicating slow-down of overall fungal metabolism. A modest number of enriched categories were in late up-regulated genes, including iron starvation, detoxification and stress. Overall, the gene expression patters and functional analysis suggested that human blood is not a hostile environment for *A. fumigatus*, which first senses the environment and then shuts down several important pathways of energy-consuming metabolism after the first hours and thus can not effectively grow in blood.

Kale et al. (Kale *et al.*, 2017) performed a time-series dual RNA-Seq analysis of two immunosuppressed mice models (one treated with chemotherapy and another one with corticosteroid) challenged with a particular *A. fumigatus* strain (Cea10) to assess how do host-pathogen interactions vary between two distinct immunocompromised states in pulmonary aspergillosis. Dual RNA-Seq yielded 16-29 mln reads, depending on the sample, from which 98% mapped to the mouse genome and the rest to the fungal counterpart. Further differential expression analysis of the host side revealed that the two immunocompomised models showed distinct patterns of gene expression response to the pathogen, showing that host response to *A. fumigatus* depends on the type of immunosupression. Functional enrichment analysis of DE genes showed enrichment in numerous immunological processes related to cytokines, chemokines, and their receptors in the chemotherapeutic model, whereas for the corticosteroid model a limited number of cytokine-related genes were DE. More highlighted differences in functionally enriched categories between the two models were found with regard to metabolic processes – the chemotherapeutic model was enriched with urea cycle, pentose phosphate metabolism, nucleotide and vitamine metabolism, while corticosteroid-treated model in inositol phosphate metabolism, fatty acid oxidation, terpenoid biosynthesis, thiamine biosynthesis, etc. Additionally, the authors identified novel genes from pathogen-sensing gene families of *Tlrs*, *Clecs*, and *Nlrs*, which had not been previously described in pulmonary aspergillosis.

On the fungal side, the comparison of gene expression profiles of *A. fumigatus* in different mice models has shown a large proportion of similar DE genes (n=3345), with a restricted number model- and time-point specific DE genes (n=128-204). Nevertheless, the analysis of fungal secreted proteins, which are important for fungal pathogenesis, showed that *A. fumigatus* has a temporal and model-specific activation of these proteins. However, it has to be noted that the analysis of a large proportion of fungal genes (5175, ~60% of total genes) was precluded due to very low expression values, which once again demonstrates the problem of low concentration of fungal genetic material in *in-vivo* dual RNA-Seq experiments.

Previously it has been demonstrated that virulence varies among different *A. fumigatus* strains (Fuller *et al.*, 2016; Kowalski *et al.*, 2016), and moreover that the host immune response against different strains is also distinct (Rizzetto *et al.*, 2013). Thus, as opposite to Kale et al., Watkins et al. (Watkins *et al.*, 2018) have recently made RNA-Seq gene expression profiling of two *A. fumigatus* strains *Cea10* and *Af293* interacting with human airways cell line A549 to elucidate the differences and commonalities between the virulence mechanisms of aforementioned strains. The experiment comprised two time-points (6 and 16 h post infection) of infected human cells with two fungal strains, and time-matched controls for fungal samples (i.e. fungi without host cells). The study focused only on the fungal side thus the authors did not perform controls for the host counterpart. Since the ratio of fungal and human cells in the infection models was close to 1:1, the RNA-Seq successfully recovered enough amount of fungal reads (53±29.3 million reads per sample) for robust downstream analysis. Differential expression analysis revealed 7,888 genes across conditions, and PCA and hierarchical clustering with those genes showed that samples clustered largely according to time-points and by strain (on 16h time-point), and that controls were positioned close to infected samples. Taken together, the patterns of sample distribution and clustering highlighted that changes of fungal transcriptional profiles are largely due to growth and metabolism dynamics, rather than to a response to the lung epithelial cells. To dissect the genes that are specifically involved in the infection process, the authors compared transcriptional profiles of time-matched infection and control samples. In this case, a modest response to human cells was found (n=128-619 DE genes) with 70% of them being strain-specific, indicating the virtual lack of strong conservative response of *A. fumigatus* strains at least in the analyzed conditions. Nevertheless, a small proportion of genes (n=47) that were similarly expressed in both strains was found, and the mutants of seven of these genes were shown to have attenuated virulence.

Chen et al. (Chen *et al.*, 2015) investigated the interactions of A549 epithelial cell line with *A. fumigatus*, but unlike in Watkins et al, this study focused on the host side. Here, the authors infected cell cultures with *A. fumigatus B5233* for 8 hours, and performed fungal-free control at similar conditions, with further RNA isolation and sequencing of the host cells. Differential expression analysis between infected and intact cells revealed in total 302 up- and 157 down-regulated genes. GO enrichment analysis showed that down-regulated genes were enriched in ion transport, skeletal system developments and vascular development. On the other hand, similar to other studies, up-regulated genes were functionally enriched in numerous immune-associated processes such as chemotaxis, inflammatory response, response to bacterium, and also in cytoskeleton remodeling, which also has been reported earlier (Jia *et al.*, 2014). To further investigate the role of specific host genes in fungal response, the authors chose two genes – *ARC* and *EGR1* – involved in cytoskeleton rearrangements, since it is known that *A. fumigatus* conidia are able to internalize in the host cells. Western blotting indicated that corresponding proteins of that genes were up-regulated during the course of infection. Moreover, inhibition of expression of *ARC* and *EGR1* genes by RNAi decreased the internalization rates of conidia by 20% and 40%, respectively.

Another *in-vivo* murine infection model of host-aspergillus interactions was made by Shankar et al (Shankar *et al.*, 2018). Here, unlike in the above mentioned study by Kale et al (Kale et al. 2017), the authors investigated invasive aspergillosis in immunocompetent mice, and focused on kidney infection. The study comprised a time-course infection model at five different time points up to eight days after infection. At all time points and for control animals, infected kidneys were homogenized and subjected to RNA-Seq. Initially the authors aimed for resolving host-pathogen interaction of both counterparts, however as usual in *in-vivo* studies, fungal side did not yield interpretable amount of data. Thus further analysis was focused only on the host transcriptome. Overall, differential expression analysis revealed more than 14,000 DE genes throughout the course of infection in mice. Although notable up-regulation was observable from the first day after infection, functional enrichment was only observed after five days post infection. Enriched terms included leukocyte aggregation, acute inflammatory responses, positive regulation of chemokines, and other immune response processes. A more in-depth investigation of up-regulated genes showed the activation of several genes such as *Ccr5*, *Cxcr3*, *Ccr2*, or *Cxcr4*, which are directly involved in activation and recruitment of Th-1 and Th-17 T-helper cells. In turn, after activation of Th cells the up-regulation of different proinflammatory cytokines such as *IFN-c*, *IL-27*, *IL-18*, *IL-24* was detected. Conversely, down-regulated genes were associated with iron and heme binding, electron carrier activity and aromatase activity. However, in the case of iron-regulation associated genes, the pattern of down-regulation was explained by suppression of P450 related genes, since many other key components of iron homeostasis, like *Nos1*, *Nos2*, *Ltf* were systematically up-regulated.

3.3.3 *Cryptococcus*

Unlike *Candida* and *Aspergillus*, the *Cryptococcus* genus belongs to the phylum *Basidiomycota* (http://www.asmscience.org/content/book/10.1128/9781555816858.ch01). There are two main pathogenic *Cryptococcus* species, *Cr. neoformans* and *Cr. gattii* which are environmental non-host-specific pathogens infecting a wide range of hosts including insects, plants and mammals. In the case of humans, *Cr. neoformans* is mainly an opportunistic pathogen, while *Cr. gatii* can infect immunocompetent individuals (reviewed in Kwon-Chung *et al.*, 2014). In recent decades, the incidence of cryptococcosis has increased drastically which is mainly associated with an emergence of HIV and increasing numbers of organ transplant recipients. The main types cryptococcal infections are cutaneous cryptococossis, pulmonary cryptococossis and cryptococcal meningitis, with the latter being fatal if not treated on initial stages. In the developing part of the world it has been estimated that these two species cause around one million infections with morality rates reaching 70% and causing 650,000 deaths per year (Park *et al.*, 2009; Brown *et al.*, 2012). As it is the case of *Aspergillus* conidia, cryptococcal spores or dried yeast cells enter the host organism through inhalation or through direct interaction in the case of skin-related infections.

As for *Candida* and *Aspergillus*, the pathogenicity mechanisms of *Cryptococcus* species have been more extensively studied *in-vitro* by exposing them to different environmental conditions (O'Meara *et al.*, 2013; Zhang, Park and Williamson, 2015; Brandão *et al.*, 2018). However, *in vivo* or *ex vivo* transcriptomic studies of *Cryptococcus*-host interaction are limited to several recent studies. Moreover, some studies using RNA-Seq were performed for refining the genome and transcriptome annotations of *Cryptococcus* species (Janbon *et al.*, 2014; Gonzalez-Hilarion *et al.*, 2016; Ferrareze *et al.*, 2017), which are not covered by our review.

The first study investigating transcriptome of *Cr. neoformans* cells interacting with the host environment was carried out in 2014 by Chen et al (Y. Chen *et al.*, 2014). The authors performed RNA-Seq of two *Cr. neoformans var. grubii* strains, G0 and HC1, taken directly from the cerebrospinal fluid (CSF) of two patients with cryptococcal meningitis. Additionally, the same fungal cells were grown in two conditions – *ex-vivo* CSF and YPD, followed by RNA sequencing and comparison with *in-vivo* obtained fungi. Initial analysis showed that gene expression profiles of both strains in each condition were very similar, thus the strains at the given conditions were considered as biological replicates. Differential gene expression analysis between pairs of conditions identified 129, 45, 256 DE genes when comparing *ex-vivo* versus YPD, *in-vivo* versus YPD and *in-vivo* versus *ex-vivo*, respectively. This shows that transcriptomes from *in-vivo* and YPD samples are more similar that in *ex-vivo* CSF samples. Compared to *ex-vivo* cells, *in-vivo* samples were enriched with cellular biosynthetic GO terms, indicating that *Cr. neoformans* cells within the host are transcriptionally more active, which might be explained by interaction with the immune system. On the other hand, as expected, samples exposed to CSF (in both cases) comparing with those with YPD had multiple DE genes previously reported to be important for *Cr. neoformans* virulence, such as such as *CFO1* (Jung *et al.*, 2009), *ENA1 (Idnurm et al., 2009)*, and *RIM101* (O'Meara *et al.*, 2010). The authors also identified 100 strain-specific differentially expressed genes, which were enriched in transporter genes. Additionally, the high sequencing depth allowed the authors to perform variant calling of sequenced strains and compare their genotypes with the reference. Variant calling showed a substantial genomic variation between the analyzed strains and the reference genome – 50,155 and 156,880 SNVs were identified in G0 and HC1, respectively, which demonstrates that diverse *Cr. neoformans* strains have largely similar transcriptomic responses to the host environment. Taking advantage of the high depth and quality of the sequencing data, the authors performed *de-novo* assembly of the transcriptomes of strains identifying novel genes related to transport, localization, and membrane constitution. Taken together, this work was the first study addressing the question of virulence mechanisms of phylogenetically diverse strains of *Cr. neoformans* obtained directly from host using RNA-Seq and thorough methods of bioinformatics data analysis.

In another study Li et al. (Liu *et al.*, 2014) compared the transcriptomic profiles of brain tissues in mice, infected by WT *Cr. neoformans* and double knock-out mutants for the genes of inositol transporters Itr1a and Itr3c, which were previously shown to be involved in fungal virulence through their role in uptaking inositol from the host. Gene expression profiles obtained by RNA-Seq were generated for control mice, and were compared in a pairwise manner with those from mice infected by the two fungal strains. Differential expression analysis identified 1133 up- and 1600 down-regulated genes in WT-infected mice, while *itr1aΔ itr3cΔ* mutant strain showed altered expression of 552 up-regulated and 278 down-regulated genes. 371 genes were shared between mice infected by each of the two strains. GO enrichment analysis showed that many enriched functional terms are shared across the two different infections, including cellular death and survival, cell-to-cell interaction, involvement in neurological disease, etc. However, mice infected by the mutant strain extensively activated immune-related responses, such as inflammation, humoral immune response, free radical scavenging, etc. In stark contrast, none of the immune response pathways was significantly enriched in WT-infected mice. Moreover terms related to cell death and necrosis were enriched only in WT-infected mice. To assess changes in the pathogen that were resulting in different host responses, the authors measured the size of fungal capsule and the secretion of glucuronoxylomannan (GXM) – two important factors for *Cr. neoformans* virulence. While the capsule size was similar in the two strains, the secretion of GXM was significantly reduced in the *itr1aΔ itr3cΔ* mutant. This result was also confirmed by immunohistologic staining of GXM in mouse brain tissue, showing that animals affected with mutant had less GXM around brain lessons. Overall, this study demonstrated the role of inositol transporters in host-pathogen interactions, linking their function with the secretion of GXM and altered composition of the fungal capsule, which in turn elicits a highlighted host immune response.

Hu et al. (Hu *et al.*, 2014) investigated the ability of environmental *Cryptococcus neoformans* strains to undergo microevolutionary changes promoting the increase of virulence during serial host passages. The authors used nine haploid serotype A *Cr. neoformans* strains isolated mainly from soil. Each strain was inoculated into mice sequentially four times over four months. Each following passage to a new mouse was performed using fungal colonies isolated from brains of the precedent mouse. Two strains were revealed with prominent increase of virulence, which in both cases have reduced the time of mice death of the first and the last passages by 4-

fold (~ 25 hours in the 1st infected mouse and ~ six hours in the 4th infected mouse). To disentangle the transcriptomic changes of highly adapted strains compared to their environmental predecessors, the authors have performed RNA-Seq of aforementioned two strains and one control strain, that did not show virulence changes across the passages. RNA-Seq analysis revealed four genes with significantly higher expression in evolved strains compared to predecessors. One of them, *Fre3* (*CNAG_06524*), was shared between two species. Using RNA interference, the authors identified that *Fre3* functions as an iron reductase without copper reductase activity. To confirm the role of the gene in pathogenicity, they over-expressed *Fre3* in WT background, which recapitulated the increased adaptive virulence phenotype. Overall, the study shows how RNA-seq can be used to address the important process of enviroment-to-mammal transition of *Cr. neoformans*, identifying the role of iron reductase *Fre3* in the adaptation to the host.

3.4 Emerging technologies in RNA-Seq

3.4.1 Single cell RNA-Seq

The term RNA-Seq is generally referred to sequencing of RNA, which was isolated from the population of cells (bulk RNA-Seq). Thus, the results obtained from bulk RNA-Seq constitute an averaged signal from the sum of individual cells, while each cell (or a sub-population of cells) might have its own transcriptomic patterns. The limitation of sequencing the bulk RNA was overcome by two major technological advances: efficient cell sorting with single cell isolation, and the availability of efficient protocols for the amplification of minute amounts of RNA from these single cells (Kolodziejczyk *et al.*, 2015). Today these two methods and their derivatives allow performing single cell RNA sequencing (scRNA-Seq), disentangling transcriptional profiles of thousands (even hundred of thousands) individual cells (Fan, Fu and Fodor, 2015; Zheng *et al.*, 2017; Rosenberg *et al.*, 2018). Despite challenges related to cost, technology and data analysis (Kolodziejczyk *et al.*, 2015; Stegle, Teichmann and Marioni, 2015; Weinreb *et al.*, 2018), scRNA-Seq is now one of the most precise methods in transcriptomics studies. However, as compared to studies on mammals, it has not been used much to study microbial cells (Kolisko *et al.*, 2014; J. Wang *et al.*, 2015; Rosenthal *et al.*, 2017) and host-pathogen interaction studies (Avraham *et al.*, 2015; Saliba *et al.*, 2016). One recent advancement in this field was reported in Avital et al., 2017 (Avital *et al.*, 2017), where the authors developed a method for single cell dual RNA-Seq for mouse macrophages and *Salmonella typhimurium* cells during infection, revealing three distinct stages of macrophage response to the pathogen. On the other hand, to our knowledge there are no studies addressing human-fungal interaction on the single cell level. This technology holds a great potential to unravel specific expression patterns governing the switches between fungal morphotypes, quorum sensing, switches from commensalism to pathogenicity and switches between the stages of infection. Moreover, single cell transcriptomics approaches can decipher how the host senses and reacts to the pathogen at different infection stages and deconvolve the expression patterns of different cell types, especially in context of *in-vivo* studies.

3.4.2 Long-read sequencing

Today, the dominating sequencing technology "sequencing-by-synthesis" of Illumina, also known as second generation sequencing, generates relatively short reads (25-300 bp) with a very high throughput and high accuracy. Despite the great advantage of the last two features, short reads are often problematic in some specific tasks, such as assembly of complex and repetitive genomes or accurate reconstruction of transcript isoforms. To overcome this problem, Illumina has recently implemented so called TruSeq synthetic long-read technology, previously known as Moleculo (McCoy *et al.*, 2014). This experimental and data analysis approach splits the molecule into smaller pieces and uses barcodes to tag the adjacent sequences. Further sequencing and bioinformatics data analysis reassembles the initial sequence, thus allowing to obtain synthetic long reads.

Moreover, in the last decade, the advent of third generation sequencing has opened new avenues in biomedical research, allowing to sequence much longer reads (up to several hundred kbs (Jain *et al.*, 2018)). Moreover, their single molecule sequencing technology is PCR free, which eliminates potential PCR amplification biases. However, today long read sequencing comes with two major disadvantages, which are low throughput and high error rates as compared to sequencing-by-synthesis technology. The two major technologies for long read sequencing are operated by Oxford Nanopore (ON) and Pacific Biosciences. The details of each technology are reviewed in (Lu, Giordano and Ning, 2016) and (Rhoads and Au, 2015), respectively. Although long-read sequencing was initially used in genomics field to assemble more contiguous and resolved genomes, today the technologies have also been validated in transcriptomics applications, mainly in transcript discovery (Sharon *et al.*, 2013; Chen *et al.*, 2017; Garalde *et al.*, 2018) and at lesser extent in gene/transcript expression profiling (Byrne *et al.*, 2017). To fill the gap of low throughput of long read technologies, and thus allow reliable expression evaluation, so-called hybrid-sequencing can be used, which utilizes both short and long-read sequencing data (Ning *et al.*, 2017; B. Wang *et al.*, 2018). The third generation sequencing has already advanced our knowledge about the transcriptomes of even well studied organisms, identifying numerous previously uncharacterized transcripts and splicing events (Au *et al.*, 2013; Sharon *et al.*, 2013; L. Chen *et al.*, 2014; Byrne *et al.*, 2017; Chen *et al.*, 2017). Moreover, it already has been reported that ON can be effectively used in microbial diagnostics (Quick *et al.*, 2015; Mitsuhashi *et al.*, 2017; Schmidt *et al.*, 2017), providing the potential of identifying the pathogen in 2-4 hours.

Overall, long read sequencing technology can dramatically further our knowledge of transcriptomes of poorly studied organisms, as in case of the most of human fungal pathogens. In this case, novel-specie-specific transcripts can become promising biomarkers for fungal diagnostics and discovery. On the other hand, when applied to host-pathogen interaction studies, it might allow the precise reconstruction of novel pathogen-specific transcripts, like lncRNAs, in the host side, which have been already shown as immune response regulators (Heward and Lindsay, 2014; Ouyang, Hu and Chen, 2016; Jiang *et al.*, 2018).

Nevertheless, the third generation sequencing is still in its infancy, and further improvements and validations of the technology are necessary in order to make it more versatile and popular in biomedical research.

3.4.3 Potential applications of RNA-Seq in fungal diagnostics

Next-generation sequencing methods have become increasingly popular in the clinics, especially in the context of diagnosis of cancers (Luthra *et al.*, 2015) and Mendelian diseases (Jamuar and Tan, 2015). Moreover, today these techniques have already penetrated to microbiology labs, allowing to achieve high precision of microorganism detection (Turabelidze *et al.*, 2013; Shaw *et al.*, 2016), identify drug resistance (Stoesser *et al.*, 2013; Wain and Mavrogiorgou, 2013), control outbreaks (Reuter *et al.*, 2013; Sherry *et al.*, 2013) and to study microorganisms that are difficult to grow using conventional culturing methods (Berenguer *et al.*, 1993). However, despite the fact that the incidence of fungal infection is steadily increasing, so far the efforts in applying NGS in microbial diagnostics have been mainly focused on bacteria and viruses. However, fungal pathogens possess features that make them difficult for management under the paradigm of traditional microbiology diagnostic methods, such as, as discussed above, rapid emergence of antymycotic drug resistance, emergence of new pathogenic species, high biodiversity, etc. Thus, the necessity of novel analytical tools, such as NGS in fungal diagnostics becomes inevitable. On the other hand, a distinction of different NGS tools in their applicability for diagnostic purposes has to be done. While DNA sequencing plays a the major role for species detection, identification and characterization, RNA-seq holds a great potential in identifying biomarkers (in form of novel transcripts) and gene/transcript expression level signatures specific to different species or for different stages of infection. Nevertheless, to achieve this kind of diagnostics, additional research efforts have to be performed. Precisely here is where the emerging technologies can immensely further the potential for RNA-Seq diagnostics. For instance, inherent problems such as the low amount of fungal RNA in patient samples can be effectively solved using probe enrichment, while the further identification of novel transcripts is addressable by long-read or hybrid sequencing. Moreover, single-cell RNA-Seq approach could be applied to decipher transcriptomic differences between cell populations, increasing the potential resolution of diagnostics. On the other hand, prices and turn-around time of these technologies are yet to achieve levels that make them suitable for the clinical settings. However, with current trends of diminishing prices, smaller and easier to handle machines, and faster turn-around times the future of RNAseq based diagnostics may be approaching. Taken together, RNA-Seq and related methods open promising avenues for fungal diagnostics, but nevertheless still a considerable research and technical developments have to be carried out to truly uncover this potential.

3.5 Concluding remarks

In the recent decade the advent of transcriptome sequencing technologies has opened exciting possibilities for exploring gene regulation and how it varies in different contexts at a level of detail and throughput that surpasses the most optimistic expectations of the previous decade. Many biological disciplines are taking advantage of this new era in transcriptomics, and host-pathogen interaction studies are no exception. As a result our knowledge about the molecular mechanisms of the interplay between various microbes and their hosts has greatly advanced in this time-frame. While the application of RNA-Seq for unraveling human-fungal interactions is just gaining momentum, it is already clear that the use of this technology and its derivatives will be the main trajectory in the field for the coming years. Despite its versatility, today RNA-Seq faces several natural and technical barriers, specifically in human-fungal interaction studies. While *in vivo* studies are complicated by extremely low amount of pathogen cells, infection models do not entirely reconstitute the whole complexity and peculiarities of human-fungal interactions. On the other hand, RNA-Seq is still relatively expensive, and requires specific expertise in study planning, bioinformatics data analysis, and interpretation of results. Moreover current mainstream technologies are limited in several technical aspects. Nevertheless, the technological advancements in the field are occurring at a fast pace and they are already partially overcoming most of the aforementioned limitations. We anticipate that dual host-pathogen RNA-Seq analyses in both *in vivo* models and patients will multiply in the coming years, as current limitations are overcome, and will constitute the basis of key advancements in our understanding of host-pathogen interactions during commensalism and infection. Finally, although there are still many technical and practical impediments for the use of RNA-Seq for diagnostic purposes, we foresee a great potential that may be realized as key biomarker genes of the process of infection are discovered and technical developments enable bringing fast, accurate and affordable RNA-Seq based technologies to the clinics.

Thesis objectives

In this PhD thesis, we applied comparative transcriptomics to address two major goals - to enhance our understanding of host-microbe interactions between human and *Candida* pathogens, and to assess the impact of hybridization on yeast transcriptomes and its potential contribution to pathogenicity emergence.

To reach these goals we posed the following objectives:

- To assess the differences and similarities of host-pathogens interactions mechanisms across the major *Candida* pathogens and human epithelial cells at the transcriptomic level;

- To define and characterize long non-coding RNAs in the major *Candida* pathogens and assess their potential role in the process of infection;

- To design and validate a pan-*Candida* probe enrichment kit, allowing to amplify fungal genetic material from clinical samples of candidiasis, facilitating *in vivo* studies of host-pathogen interaction by transcriptome sequencing;

- Study the pattern of allele-specific gene expression in the hybrid fungal pathogen *C. orthopsilosis*, assessing the impact of hybridization in virulence emergence in this yeast;

- To systematically assess the effect of inter-specific hybridization on transcriptomic profiles of parental species and corresponding hybrids in yeasts;

- To develop computational tools facilitating the planning of experimental designs for studies similar to the ones performed in this thesis, which involve sequencing multiple organisms at once.

4 Diverse *Candida* pathogens induce protective mitochondria-associated type I interferon signalling and a damage-driven response in epithelial cells

Marina Pekmezovic*, Hrant Hovhannisyan*, Elise Iracane, João Oliveira-Pacheco, Eric Seemann, Britta Qualmann, Selene Mogavero, Mark S. Gresnigt, Sascha Brunke, Geraldine Butler, Toni Gabaldón, Bernhard Hube. *"Diverse Candida pathogens induce protective mitochondria-associated type I interferon signalling and a damage-driven response in epithelial cells" Nature Microbiology* (*in revision*)

* - These authors share equal first authorship. M.P. performed all the laboratory experiments (except TEM), analyzed the data, wrote the manuscript and prepared the figures. H.H. performed all bioinformatic analyses, wrote the manuscript and prepared the figures.

4.1 Abstract

Vaginal candidiasis is an extremely common disease predominantly caused by four phylogenetically diverse species: *Candida albicans, C. glabrata, C. parapsilosis*, and *C. tropicalis*. Here we show that these species exhibit distinct pathogenicity patterns, indicated by highly species-specific transcriptional profiles during infection of vaginal epithelial cells. In contrast, host cells exhibit a homogeneous response to all species at early stages of infections, which is characterized by sub-lethal mitochondrial signalling inducing a protective type I interferon response. At later stages, the transcriptional response of the host diverges in a species-dependent manner. This is primarily driven by the extent of epithelial damage elicited by species-specific mechanisms such as secretion of the toxin candidalysin by *C. albicans*. Our results uncover a dynamic, biphasic response of vaginal epithelial cells to *Candida* species and a protective type I interferon response induced by sub-lethal mitochondrial signalling, which is correlated to conserved fungal triggers and species-specific damage potential.

4.2 Introduction

The incidence of fungal infections has steadily increased over the last decades (Bitar *et al.*, 2014; Guinea, 2014; Cortegiani *et al.*, 2018). These range from superficial infections of skin or mucosa, which affect approximately 25% of the global population (Havlickova, Czaika and Friedrich, 2008), to life-threatening invasive mycoses (Klingspor *et al.*, 2015), which annually kill approximately 1.5 million people worldwide (Brown *et al.*, 2012).

Vulvovaginal candidiasis (VVC) is among the most common fungal infections, affecting 70-75% of women at least once in their lifetime (Mårdh *et al.*, 2002). VVC is characterized by acute inflammation of the vulva and vaginal mucosa due to the overgrowth of *Candida* spp. that exist typically as commensals of the mucosa (Fidel *et al.*, 2004; Yano *et al.*, 2019; Rosati *et al.*, 2020). Although *C. albicans* is the most predominant etiologic agent of VVC, the prevalence of other species like *C. glabrata, C. parapsilosis,* and *C. tropicalis* has increased (reviewed in Makanjuola, Bongomin and Fayemiwo, 2018). Despite their shared genus name, these species are phylogenetically diverse, and they often have non-pathogenic close relatives, which indicates that their ability to infect humans may have emerged independently (Gabaldón, Naranjo-Ortíz and Marcet-Houben, 2016). However, the different and likely independently evolved pathogenic interactions of these diverse *Candida* species with host cells have rarely been addressed on a comparative basis. Clearly, a better understanding of the similarities and specific characteristics of the pathogenic processes during infection with each species would be crucial to improve diagnostics and therapy of *Candida* infections (Consortium OPATHY and Gabaldón, 2019).

Neutrophils and macrophages are considered to be the crucial players in the first line of defence against fungal infections in mammals, followed by adaptive T-cell immunity, and thus research on host-fungal interactions has been mainly focused on these cell types (Verma, Gaffen and Swidergall, 2017; Meir *et al.*, 2018). However, a fundamental role is reserved for epithelial cells in initiating and shaping the host defence against fungi which goes far beyond their function as a physical barrier (Moreno-Ruiz *et al.*, 2009; Moyes *et al.*, 2010; Zhu and Filler, 2010; Moyes and Naglik, 2011; Naglik and Moyes, 2011; Naglik *et al.*, 2011).

Although the overwhelming majority of studies in infection biology focus on either the pathogen or the host immune response, microbial pathogenesis can be best interpreted in the framework of dynamic host-microbe interactions. Recent developments in next-generation sequencing such as dual RNA-sequencing (RNA-Seq) enable combined assessment of the transcriptional response of host and pathogen throughout the course of infection (Hovhannisyan and Gabaldón, 2019). This approach has been successfully used to address the interactions of fungal pathogens with different host cells (Tierney *et al.*, 2012; Amorim-Vaz *et al.*, 2015; Bruno *et al.*, 2015; Liu *et al.*, 2015; Tóth *et al.*, 2018), although primarily focusing on *C. albicans.*

To elucidate the general and species-specific molecular patterns of interactions between vaginal epithelial cells and the four most-prevalent VVC-causing *Candida* species, we used a time-course *in vitro* infection model coupled to dual RNA-Seq. Our experimental design allows the pathogens to deploy their whole arsenal of pathogenicity mechanisms without being restricted by the host's innate or adaptive immune system. Further, it facilitates specifically investigating the recognition and defence mechanisms employed by epithelial cells, which constitute the first line of defence against infecting fungi.

Our results reveal that fungal transcriptomes show species-specific patterns, likely reflecting the independently evolved pathogenic potential of the most common *Candida* species. In contrast, human epithelial cells display a biphasic transcriptional response, including a conserved early response, which is partially independent of fungal viability. Importantly, we identify a viability-dependent, protective type I interferon (IFN) response, mediated by sub-lethal mitochondrial signalling, and a damage-associated late response depending on species-specific pathogenicity mechanisms.

4.3 Results

Candida species-specific pathogenicity patterns

We focused on the interaction of the four most common *Candida* species causing VVC with vaginal epithelial cells. First, we assessed the adhesion, invasion, and damage potential as well as the morphology of all four species in an *in vitro* vaginal epithelial infection model (Fig. 4.1). Despite similar adhesion rates of all species (6.7-10.7 %), only *C. albicans* switched to true hyphal growth, invaded the epithelial cells, and inflicted a high level of damage. Non-invading *C. glabrata* and *C. tropicalis* cells caused similar intermediate damage levels. *C. glabrata* cells grew only in the yeast morphology, whereas occasional pseudohyphae were observed for *C. tropicalis*. Finally, *C. parapsilosis* remained in the yeast morphology during the entire course of infection, forming cell aggregates, but did not invade or damage epithelial cells. These results show species-specific pathogenic patterns, with an initial uniform attachment efficiency followed by diversification, involving different morphologies, levels of invasion, and damage capacity.

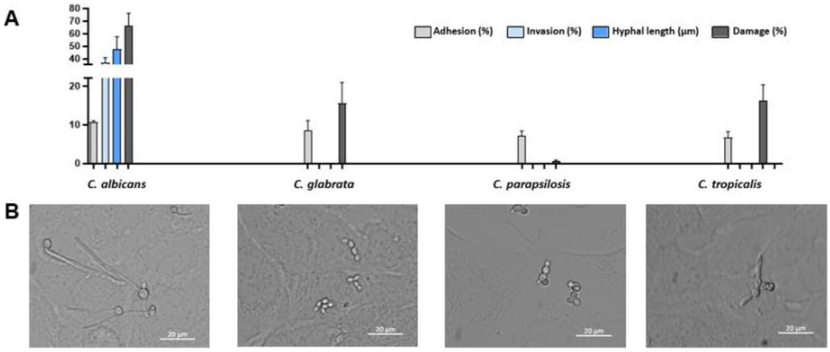


Fig. 4.1. Pathogenic patterns of four *Candida* species in the *in vitro* vaginal epithelial infection model. (A) Adhesion (%; 1 hour post-infection - hpi), invasion (%; 3 hpi), hyphal length (µm; 3 hpi), and damage (% of the maximum damage control; 24 hpi) of *Candida* species on vaginal epithelial cells (ECs). **(B)** Micrographs of *Candida* morphology on vaginal ECs at 3 hpi. All values are presented as mean ± SD, n=3.

Dual RNA-seq of Candida - epithelial cell infection model

Subsequently, we investigated whether these differential pathogenic patterns also differ on a transcriptional level for both fungal and epithelial cells by performing dual RNA-seq at several time points chosen according to the observed infection events (0, 1.5, 3, 12, and 24h).

We assessed potential cross-mapping between reads from human and the different *Candida* species using Crossmapper (Hovhannisyan, Hafez, *et al.*, 2020), and designed a pooling and sequencing strategy that resulted in virtually no cross-mapping (suppl. files S1-S4). Overall, our experiment yielded ~4.5 billion paired-end reads for 96 samples for the different time points, infecting species, and controls, of which ~2.3 and ~1.7 billion reads mapped uniquely to yeast and to host reference genomes, respectively. This constitutes one of the largest and most comprehensive RNA-Seq datasets of human-*Candida* interactions. An overall description and summary statistics of our dataset can be found in suppl. table S1. Mapped data from the four *Candida* species can be mined and browsed at Candidamine (candidamine.org).

Species-specific transcriptional responses to epithelial cells

We first analysed the transcriptional dynamics of each *Candida* species throughout the infection (Fig. 4.2). The complete set of differential gene expression data can be found in suppl. files S5-S8. The four *Candida* species have a rapid transcriptional response following infection (Fig. 4.2A). A large number of genes (405-919) were differentially expressed in all species at 3 hours post-infection (hpi) as compared to baseline expression. When comparing differentially expressed genes across species, a remarkably distinct pattern was observed for each pathogen (Fig. 4.2B-D).

First, we assessed the distribution of species-specific (i.e. genes without orthologs in the other three species), partially shared (i.e. a gene in a given species with orthologs in 1 or 2 of the other species) and fully shared genes (i.e. a gene in a given species has orthologs in all other species) with differential expression during infection (Fig. 4.2B). Species-specific and partially shared genes constitute a substantial proportion (31-72%) of differentially expressed genes in each of the four *Candida* species at any time point. Moreover, species-specific genes overall are more likely to be differentially expressed than fully shared genes (Chi-square test $p<0.05$, except for *C. tropicalis*). Furthermore, even fully shared orthologous genes showed species-specific differential expression (Fig. 4.2C), further stressing the presence of distinct molecular patterns of host-pathogen interactions. Consistently, Principal Component Analysis (PCA) based on gene expression levels showed species-specific clusters, independently of the time point of infection (Fig. 4.2D).

Our infection model involves a change of growth medium for the yeasts. To account for this and to identify genes that are specifically expressed in response to epithelial cells ("infection-specific genes"), we included time-matched controls in culture medium only (suppl. Fig. S4.1). We subtracted the list of differentially expressed genes observed at 24-hour control (hc) in fungal-only controls from those found at 24 hpi (see suppl. Fig. S4.2 and suppl. files S5-S8, sheets "subset_0vs24"). The distribution and orthology relationships (suppl. Fig. S4.2B) of infection-specific genes (i.e. DE exclusively during infection) showed that they are mostly species-specific and partially shared.

In summary, we observed striking distinct transcriptional patterns for all fungal species supporting the view that the strategies of each *Candida* species to cope with epithelial cells during infection are specific and likely the result of independent evolution.

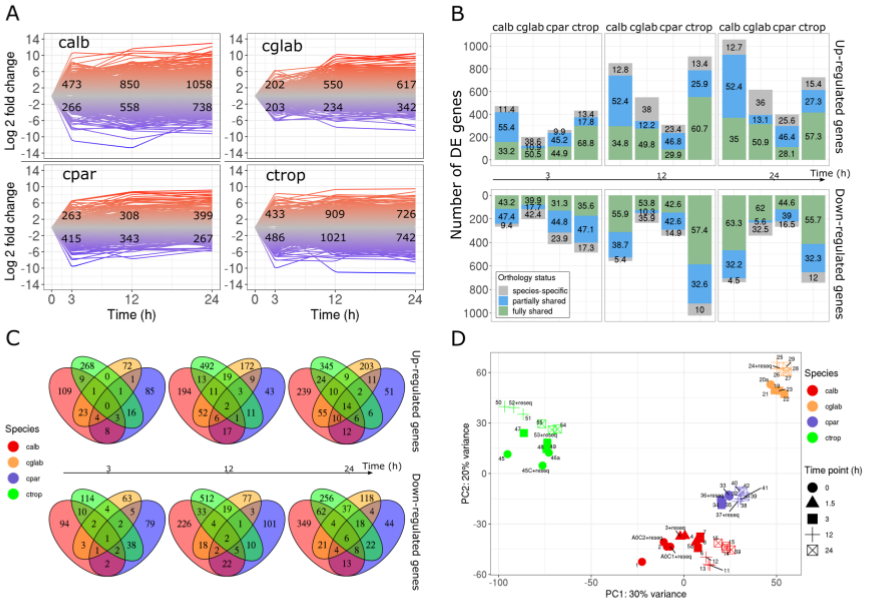


Fig. 4.2. Dynamics of transcriptomic changes of the four *Candida* species investigated in this study at different time points. (A) Transcriptome dynamics plots based on log2 fold changes compared to time point 0. Each line corresponds to expression levels of a gene. Numbers on the plots indicate a number of differentially expressed (DE) genes (up-regulated - red, down-regulated - blue). **(B)** Distribution of fully shared, partially shared and species-specific DE genes across the course of infection. Numbers on bar plots indicate the percentage (%). **(C)** Venn diagrams of DE genes in four *Candida* species at each time point. **(D)** PCA biplot based on normalized expression levels of orthologous genes across *Candida* species, demonstrating a clear species-specific stratification of transcriptomic profiles of the four fungal pathogens. Labels of the data points correspond to sample IDs; "calb" denotes *C. albicans*, "cglab" - *C. glabrata*, "cpar" - *C. parapsilosis* and "ctrop" - *C. tropicalis.*

Epithelial transcriptomic responses to Candida species

Our data above indicate both distinct fungal pathogenicity and transcriptional pattern for each *Candida* species. To also shed light on the host response during epithelial-*Candida* interactions, we first assessed the overall magnitude of epithelial transcriptome changes upon exposure to the four yeast species at different time points of the infection (Fig. 4.3A and B).

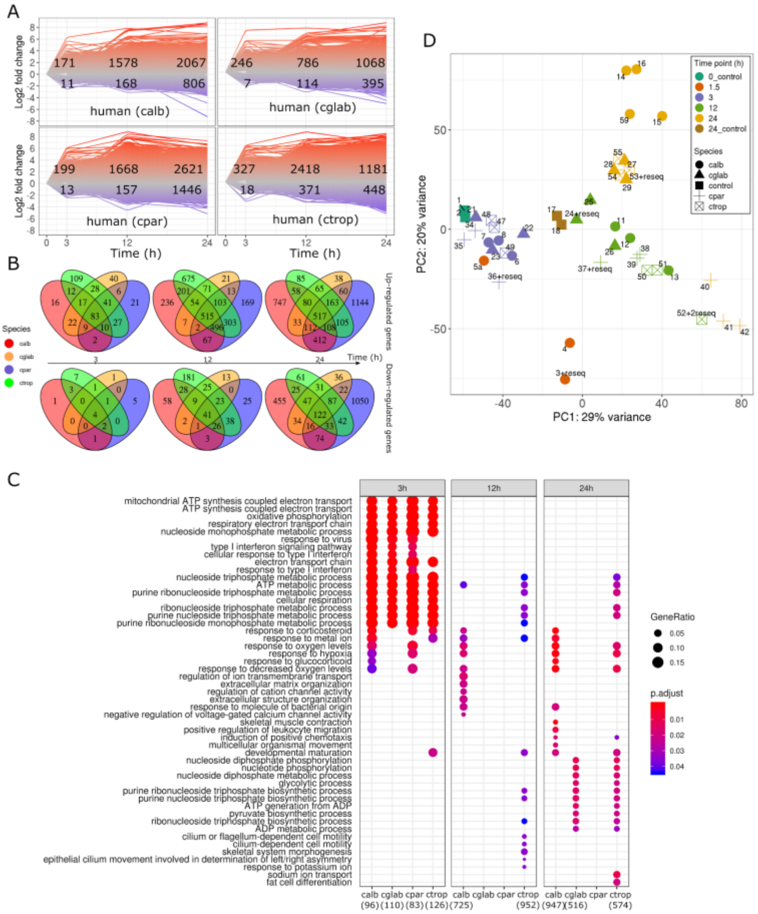


Fig. 4.3. Transcriptome dynamics of vaginal epithelial cells upon exposure to four *Candida* species. **(A)** Transcriptome dynamics plots based on log2fold changes compared to time point 0. Each line corresponds to expression levels of a gene. Numbers on the plots indicate a number of DE genes (up-regulated - red, down-regulated - blue). **(B)** Venn diagrams showing similarities and differences of human DE genes in response to *Candida* species. **(C)** GO term enrichment analysis for up-regulated genes (category "Biological Process") of the host at different time points. The x-axis indicates the infecting *Candida* species. Only significant (p_{adj}<0.05) GO enrichments are shown. Numbers in brackets underneath the species labels correspond to "counts" of the clusterProfiler Bioconductor package, i.e. the total number of genes associated with GO categories. GeneRatio corresponds to the ratio between the number of genes enriched in a given category and "counts". Adjustment of p-values is done by Benjamini-Hochberg procedure. **(D)** PCA biplot of all analysed human samples. Labels of the data points correspond to sample IDs. "calb" denotes human samples interacting with *C. albicans*, "cglab" - with *C. glabrata*, "cpar" - with *C. parapsilosis* and "ctrop" - with *C. tropicalis*.

The epithelial transcriptome dynamics indicate a clear bias towards gene up-regulation at the initial stages of infection as compared to down-regulation (Fig. 4.3A), which is consistent with the previous findings for vaginal cells in response to *C. albicans* infection (Richardson *et al.*, 2018). The number of down-regulated genes was, in some cases, 35-fold less than up-regulated genes. For example, we observed 246 up- and 7 down-regulated genes in epithelial cells upon exposure to *C. glabrata* at 3 hpi, and a similar trend was found for the other species.

We assessed to which extent the host response is common or specific to each infecting *Candida* species. The large fraction of shared differentially regulated genes (Fig. 4.3B) suggests that the response to the different yeast species is generally conserved at early infection stages (3 hpi). In contrast, groups of differentially regulated gene diverged at later stages, with a more notable split between the responses to *C. albicans* and *C. parapsilosis* infections at 24 hpi. Further functional analysis by GO term enrichment showed a similar trend: a set of enriched terms is shared across interactions with the four species at the initial time point and species-specific terms appear at later stages (Fig. 4.3C).

A similar pattern was observed by PCA (Fig. 4.3D). In stark contrast to the PCA plot based on the fungal transcriptional profiles, PCA of the gene expression of epithelial cells reveal that all epithelial samples from 3 hpi form a tight cluster, indicating a uniform transcriptional response to the four *Candida* species. Thus, epithelial cells initiate conserved responses to phylogenetically distinct *Candida* species at early stages of infection. However, the epithelial transcriptomes diverge from 12 hpi onwards depending on the infecting *Candida* species, and at 24 hpi we observed separate clusters of transcriptional responses to the different species. The transcriptional response of epithelial cells to *C. albicans* fell into one cluster, the response to *C. parapsilosis* into a second cluster and the response to *C. tropicalis* or *C. glabrata* in a third.

Based on these results, we further dedicated our study to unravel the basis of the two observed key phenomena: (1) the initial uniform host transcriptional response to four distinct fungal pathogens; and (2) the divergence of the host transcriptome response at later stages of infection.

Viable and non-viable Candida cells induce similar early responses

We sought to elucidate whether the uniform transcriptional response of vaginal epithelial cells towards phylogenetically diverse *Candida* species is driven by convergent fungal virulence programs and activities or based on recognition of pathogen-associated molecular patterns shared by all four species. Therefore, we repeated the infection experiment and transcriptional analysis with UV-killed pathogens at 3 hpi and integrated these data with our initial results (suppl. Fig. S4.5, magenta data points). PCA showed that the transcriptional response of epithelial cells to non-viable and viable *Candida* cells clustered together for every species. However, we also noted some differences. For example, the proportion of genes down-regulated in response to killed cells was much higher (except for *C. tropicalis*) than in the response to viable cells (on average 152 for non-viable cells vs 12 for viable, supplementary files S5-S8).

Host mitochondria drive the uniform early responses to Candida infections

Epithelial cells responded similarly to all four *Candida* species at early time points (Fig. 4.3). Although these responses were mostly independent of fungal viability, a subset of the conserved gene expression patterns required viable fungal cells. In particular, we observed activation of genes associated with type I interferon (IFN) signalling pathways, also termed Interferon-Stimulated Genes (ISGs) (Pervolaraki *et al.*, 2018), such as *IFI6, RSAD2, OASL, IRF9, MX2, ISG15, IFI44*, and *ZBP1*, upon exposure of epithelial cells to viable, but not to UV-killed, fungal cells ($\log_2$ fold change>1.5 for *C. albicans, C. glabrata,* and *C. parapsilosis*; $\log_2$ fold change >1.3 for infections with *C. tropicalis*). Since type I IFN responses are mostly implicated in response to viral infections, additional metagenomic analyses were performed on all reads that did not map to human or *Candida* genomes by mapping them to the entire NCBI nt database to exclude potential viral contaminations. No indications of viral genome sequences were found (data not shown). Therefore, we questioned which mechanisms may have induced type I IFN signalling pathways.

Recent studies have identified an emerging role of host mitochondria as hubs of the innate immune response, controlling various major immune pathways (Mills, Kelly and O'Neill, 2017; Mohanty, Tiwari-Pandey and Pandey, 2019). In particular, it has been demonstrated that upon the interaction of host cells with bacteria and viruses, altered host mitochondria release mitochondrial DNA (mtDNA) into the cytosol, which can act as a damage-associated molecular pattern (DAMP) (Zhang *et al.,* 2010) and activate type I IFN responses (West *et al.*, 2015). In fact, our data show that all four *Candida* species caused up-regulation of mitochondria-related genes in vaginal epithelial cells, suggesting that mitochondria-associated processes are triggered (Fig. 4.4).

Since all *Candida* species were in close contact with epithelial cells at these stages, this induction may depend on physical interactions. Indeed, when the contact of epithelial cells to *Candida* cells was prevented using a transwell system, we did not observe up-regulation of mitochondria-encoded genes (Fig. 4.4A). This shows that the observed phenomenon is both viability- and contact-dependent.

To investigate if mitochondrial DAMPs, such as mtDNA, are released in our model, we measured mtDNA levels in the cytosol of infected epithelial cells 6 hpi. Compared with an uninfected control, mtDNA concentrations in the cytosol of epithelial cells increased during the infection with each of the four *Candida* species, but not with UV-killed *Candida* cells (as quantified for *C. albicans*) (Fig. 4.4B).

In addition to the release of mtDNA, mitochondria are also known to produce mitochondrial reactive oxygen species (mtROS), which recently also emerged as critical players in the regulation of immune signalling pathways. Thus, we measured epithelial mtROS levels during infection and detected a 25% increase in mtROS production at 1 hpi with all four *Candida* species, compared to uninfected controls (Fig. 4.4C; suppl. Fig. S4.3A).

Since the production of mtROS indicates changes in the mitochondrial membrane potential (ΔΨm), we measured ΔΨm at several time points of infection (1, 3, 6, and 24 hpi). Epithelial cells infected with any of the four species showed a transient increase of ΔΨm (hyperpolarization) at 1 hpi (Fig. 4.4D; suppl. Fig. S4.3B). Later, during the course of infection, ΔΨm values decreased towards the value of uninfected control for all *Candida* infections, except for infections with *C. albicans*. In the latter case, epithelial ΔΨm increased 4-fold until 24 hpi, which possibly correlated with the specific characteristics of *C. albicans* infections.

Since these data suggest at least transient abnormality in mitochondrial function in vaginal epithelial cells during all *Candida* infections, we proposed that these modifications should be visible at a microscopic level. In fact, we observed mitochondria with altered morphologies within infected epithelial cells at 1 hpi using Transmission Electron Microscopy (Fig. 4.4E).

We noted that some mitochondria of infected epithelial cells had circular shape and/or loss of integrity, compared to intact mitochondria in uninfected epithelial cells (Fig. 4.4E). Additionally, we saw that endoplasmic reticulum regions were surrounding altered mitochondria, indicating mitophagy.

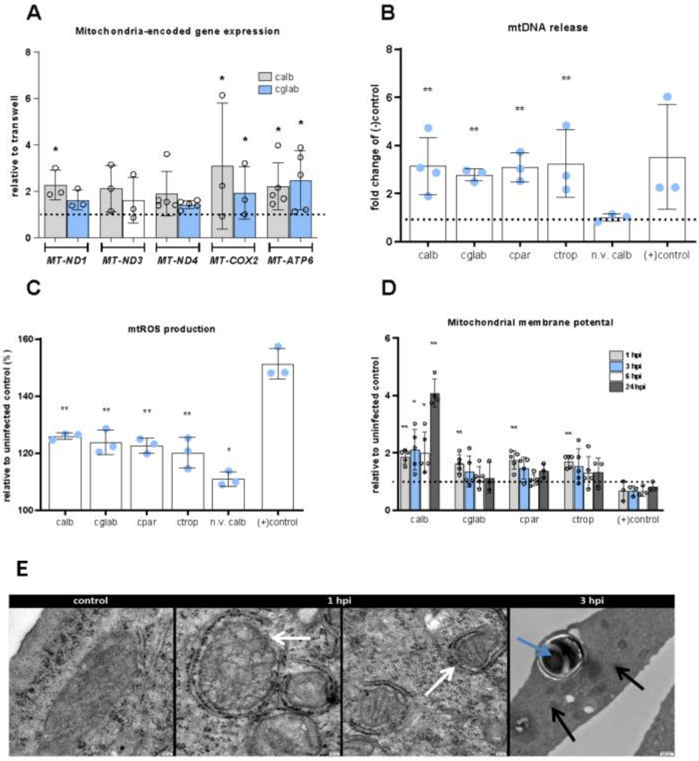


Fig. 4.4. ***Candida* species induce mitochondrial responses in vaginal epithelial cells.** Levels of **(A)** human *MT-ND1*, *MT-ND3*, *MT-ND4*, *MT-COX2* and *MT-ATP6* gene expression changes upon direct contact with *C. albicans* or *C. glabrata* compared with transwell setup. **(B)** Levels of mtDNA released into cytosol upon infection with *Candida* species or corresponding controls **(C)** Levels of mtROS production. **(D)** Mitochondrial membrane potential change (negative control: uninfected human cells; positive controls: tunicamycin 10 µM (mtDNA release); antimycin 100 µM (mtROS) and CCCP 100 µM (mitochondrial membrane potential). **(E)** Transmission Electron Microscopy (TEM) analysis of mitochondria in uninfected and *C. albicans* infected epithelial cells (1 and 3 hpi). White arrows show the loss of mitochondrial integrity in infected cells. Black arrows show circular mitochondria localizing around the invading hyphae of *C. albicans* (blue arrow) at 3 hpi. All values are presented as mean ± SD relative to the uninfected (-) control, n≥3, for TEM n=1. Statistically significantly different values are indicated by asterisks as follows: *, $p \leq 0.05$; **, $p \leq 0.01$; "calb" denotes *C. albicans*, "cglab" - *C. glabrata*, "cpar" - *C. parapsilosis*, "ctrop" - *C. tropicalis,* and "n.v.calb" – non-viable *C. albicans.*

These morphologies were observed in epithelial cells infected with any of the four *Candida* species and only occasionally in uninfected controls. Interestingly, circular shape mitochondria were also found to localize frequently around the invading hyphae of *C. albicans* at 3 hpi (Fig. 4.4E).

These all may indicate higher energy demands triggered by the presence of fungi, which lead to mitochondrial hyperactivation, visible both at transcriptional and biochemical levels, and by the production of mitochondrial DAMPs. As a consequence of affected mitochondrial function, mitophagy is potentially and selectively induced to remove damaged mitochondria in host cells.

Although mitochondrial dysfunction is a hallmark of cellular apoptosis, we did not observe any indication of epithelial cell death at early time points for any species (Fig. 4.5A; suppl. Fig. S4.4). This was expected since the observed affected mitochondrial functions were only transient and did not result in measurable damage of the host cells at that stage. Furthermore, $\Delta\Psi$m was increased rather indicating hyperpolarization, while mitochondrial depolarization would be expected in cells undergoing apoptosis. Moreover, at late stages of infection, we observed necrotic cell death, at an extent corresponding to the level of damage, and the proportion of occasional apoptotic cells was not different compared to in the uninfected control (Fig. 4.5A and S4.4).

However, to exclude that the observed mitochondrial signalling and ISG induction were associated with apoptosis, we induced apoptosis with staurosporine, infected the epithelial cells and measured the expression of ISGs at 3 hpi. As expected, we saw no upregulation of ISGs once apoptosis was induced (Fig. 4.5B).

To further verify the connection between the observed altered mitochondrial function and expression of ISGs, cytosolic mtDNA was isolated from infected epithelial cells and used to transfect uninfected epithelial cells. After 6 h, we observed up-regulation of ISGs, confirming that mtDNA released in the cytosol during the infection is, at least partially, responsible for the induction of a type I IFN response at the transcriptional level (Fig. 4.5C). Taken together, this suggests a sub-lethal engagement of mitochondrial signalling (Fig. 4.5).

Finally, we investigated the role of epithelial cells in the initiation of a pro-inflammatory response. Epithelial cells did not produce any IL-6 or IL-1β during *Candida* infection as quantified by ELISA (Fig. 4.6A). However, IL-1α and IL-8 production was detected in supernatants of *C. albicans* infected epithelial cells (Fig. 4.6A). This indicates that these tested pro-inflammatory cytokines probably do not play a significant role in subsequent immune system activation.

To ascertain whether some other pro-inflammatory molecules secreted by epithelial cells could play a role in neutrophil recruitment and activation, a hallmark of vaginal infections, we incubated primary human neutrophils with supernatants of epithelial cells individually infected with each *Candida* species (24 hpi). Control supernatants of *Candida* cells growing in media without host cells were included to ensure that molecules produced by epithelial cells (rather than fungus-derived) trigger the neutrophil response.

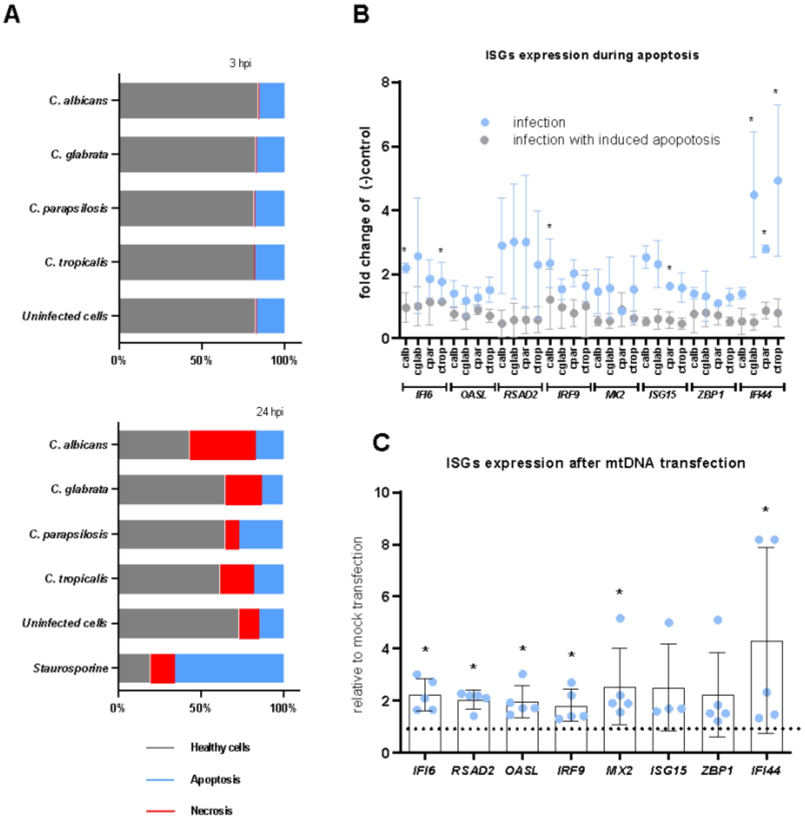


Fig. 4.5. mtDNA induces type I interferon signalling independently of apoptosis. (A) The proportion of healthy-necrotic-apoptotic epithelial cells (ECs) upon *Candida* infection at 3 and 24 hpi. **(B)** Relative expression of selected Interferon-Stimulated Genes (ISGs) (*IFI6, RSAD2*, *OASL*, *IRF9*, *MX2*, *ISG15*, *ZBP1* and *IFI44*) in infected ECs where apoptosis was induced with 1.2 µM staurosporine. **(C)** Relative expression of selected ISGs (*IFI6*, *RSAD2*, *OASL*, *IRF9*, *MX2*, *ISG15*, *ZBP1* and *IFI44*) in ECs transfected with cytosolic (mt)DNA obtained from vaginal ECs infected with *C. albicans*. Transfection with cytosolic DNA obtained from uninfected cells was used as mock transfection. All values are presented as mean ± SD relative to the uninfected/non-transfected (-) control, n≥3. Statistically significantly different values are indicated by asterisks as follows: *, $p \leq 0.05$; **, $p \leq 0.01$; “calb” denotes *C. albicans*, “cglab” - *C. glabrata*, “cpar” - *C. parapsilosis* and “ctrop” - *C. tropicalis*.

Neutrophils produced high levels of IL-8 upon the exposure to any infection supernatant (Fig. 4.6). In contrast, levels of IL-1β, IL-6 and IL-1α were below detection limits (<30 pg ml$_{-1}$).

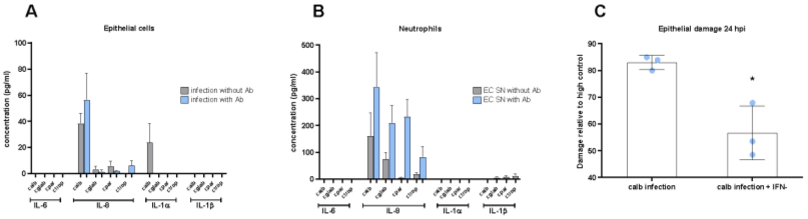


Fig. 4.6. Immune response activation and protection. (A) Levels of IL-6, IL-8, IL-1β and IL-1α secretion by ECs infected by *Candida* species 24 hpi with or without the addition of an anti-interferon-α/β receptor (IFNAR) antibody (Ab). **(B)** Levels of IL-6, IL-8, IL-1β and IL-1α secretion by neutrophils after 24 h incubation with supernatants of EC infections with or without the addition of an anti-IFNAR Ab (values obtained for incubation with fungi-only supernatants are subtracted). **(C)** Epithelial damage caused by *C. albicans* (calb) 24 hpi without and with 0.1 ng/ml of interferon-β (IFN-β). All values are presented as mean ± SD, n≥3. Statistically significantly different values are indicated by asterisks as follows: *, $p \leq 0.05$; **, $p \leq 0.01$.

Interestingly, once we infected epithelial cells in the presence of an anti-interferon-α/β receptor (IFNAR) antibody and subsequently incubated neutrophils with supernatants from infected, anti-IFNAR-treated epithelial cells, we observed a significant increase in IL-8 levels. This suggests that blocking of type I IFN signalling in epithelial cells leads to an increased pro-inflammatory response and neutrophil activation. We speculated that induction of type I IFN signalling might result in the secretion of specific anti-inflammatory molecule(s) that would prevent over-stimulation of neutrophils. Furthermore, we postulated that type I IFN signalling might increase epithelial resistance to fungal infection. Therefore, we tested whether supplementation of exogenous IFNβ may influence the level of epithelial damage caused by *C. albicans*. We observed that the presence of IFNβ (0.1 ng/ml) during the infection resulted in reduced LDH release (Fig. 4.6C), indicating that type I IFN may play a protective role during VVC.

Altogether, our data show that, at initial stages of infection, all four *Candida* species induce a similar response in vaginal epithelial cells, which involves non-lethal mitochondrial signalling *via* mtDNA release into the cytosol and mtROS production. We conclude that mtDNA released in cytosol and mtROS act as DAMPs, and initiate a type I IFN response, which has an impact on the level of neutrophil activation but also induce a protective epithelial response.

Differential cell damage drives transcriptional responses

While the initial phases of infection showed a conserved epithelial response to all four *Candida* species, this response separated into different trajectories at later stages. In particular, the response towards *C. albicans* and *C. parapsilosis* was highly divergent, while the response to *C. glabrata* and *C. tropicalis* was largely similar. We thus questioned whether defined factors or activities could explain the divergent epithelial transcriptome responses at later time points.

Considering that damage to host cells exerted by infecting microbes is one of the major determinants of pathogenicity (Casadevall and Pirofski, 2003; Pirofski and Casadevall, 2008; Jabra-Rizk *et al.*, 2016), we hypothesized that differences in the damaging potential between the infecting *Candida* species, and the subsequent response to this damage, could drive the observed differences at late infection stages. This view was supported by the fact that the damage pattern caused by *Candida* species (Fig. 4.1) showed a remarkably similar pattern to the pattern observed on the PCA plot (Fig. 4.3D): the highest level of epithelial damage was caused by *C. albicans*, followed by comparable intermediate damage by *C. glabrata* and *C. tropicalis*, which both damage at similar levels, while almost no epithelial damage was observed for *C. parapsilosis* (suppl. Fig. S4.5A).

Epithelial cell damage occurs at the final stage of *C. albicans* infection as a result of the effects of the fungal toxin candidalysin, that permeabilizes host cell membranes, leading to cell lysis (Moyes *et al.*, 2016; Wilson, Naglik and Hube, 2016). *C. albicans* mutant strains lacking the gene *ECE1*, encoding candidalysin, are almost unable to inflict damage to epithelial cells, despite normal growth, filamentation and invasion properties (Moyes *et al.*, 2016; Richardson *et al.*, 2018).
We hypothesized that candidalysin-driven epithelial cell lysis might significantly influence the transcriptional pattern of epithelial cells. To test the hypothesis, we investigated the human transcriptional response at 24 hpi upon interaction with an *ece1Δ/Δ* mutant of *C. albicans*.

Our results (suppl. Fig. S5) revealed that epithelial cells interacting with *ECE1*-deficient *C. albicans* cells at 24 hpi showed a transcriptional response notably similar to the pattern observed during *C. parapsilosis* infection, which confirms the pivotal role of damage as a major driving force of epithelial transcriptional response during fungal infections. GO term enrichment analysis for 774 genes, specifically up-regulated upon the exposure to damaging *C. albicans* wildtype, are not significantly enriched in any process. However, a recent study identified candidalysin as a key factor of *C. albicans*, which induces c-Fos and mitogen-activated protein kinase (MAPK) signaling in vaginal epithelial cells and the release of proinflammatory cytokines such as IL-1α, IL-1β, and the neutrophil chemokine IL-8 (Moyes *et al.*, 2010; Richardson *et al.*, 2018). By manual inspection, we observed similar responses on the transcriptional level (including up-regulation of *HBEGF, CXCL1, CXCL2, IL1A, IL1B, CXCL8,* and *CSF2* as well as genes associated with the "danger"-response pathway (*FOS, JUN* and *DUSP1*) in our experimental setting (Moyes *et al.*, 2010; Richardson *et al.*, 2018). Consistently, our data confirmed that epithelial damage and proinflammatory signals capable of driving neutrophil recruitment caused by *C. albicans* results almost exclusively from the action of candidalysin.

4.4 Discussion

Here we dissected the interaction of the four main *Candida* species that cause VVC (namely *C. albicans*, *C. glabrata*, *C. parapsilosis*, and *C. tropicalis*) with human vaginal epithelial cells on a cellular and molecular level. Large-scale dual transcriptomic analysis of human and fungal cells at several time points during the course of infection provided evidence for individual and species-specific *Candida* pathogenicity patterns. We observed a conserved biphasic host response with mitochondria-induced type I IFN signalling as a common early response to *Candida* infections. Only at later stages, the transcriptional responses diverged depending on the species-specific capacities to inflict damage to the vaginal epithelial cells.

We assessed how host-pathogen interactions vary across the four *Candida* species. Our results show that each tested *Candida* species has a distinct transcriptional profile upon interaction with epithelial cells. Starting from the early time point of 3 hpi, the studied pathogens showed species-specific transcriptional landscapes, even when only shared genes showing 1-to-1 orthology relationships across all four species are considered (Fig. 4.2). It has been previously hypothesized that these phylogenetically diverse species of the Saccharomycotina subphylum tree independently acquired the ability to colonize and infect humans and thus are expected to use distinct sets of molecular mechanisms during infection (Gabaldón, Naranjo-Ortíz and Marcet-Houben, 2016). Our study empirically supports this hypothesis by showing that the main VVC pathogens possess unique and species-specific transcriptional responses and patterns of pathogenicity upon contact with vaginal epithelial cells, representing the primary interaction site in the human host.

Although a uniform pattern of epithelial host transcriptional responses dominated the initial phases of *Candida* infections, this was followed at later infection time points by diverging transcriptional profiles, which varied depending on the infecting species. Notably, these diverse patterns paralleled the varying damaging capacities of the four *Candida* species. Thus, we posited that the human transcriptome divergence was primarily driven by the extent of damage induced by each species. This was confirmed by challenging epithelial cells with the non-damaging *ece1Δ/Δ* mutant of *C. albicans*. Despite a reduced damaging capacity, this mutant exhibits normal growth and filamentation. The transcriptional response to this mutant did not resemble the response to wildtype *C. albicans*. Instead, the host transcriptional response to this mutant resembled the response to the non-damaging species *C. parapsilosis*, highlighting fungal-induced damage as the major driving force of host response at late stages of infection. Furthermore, this finding confirms the crucial role of candidalysin during interaction of *C. albicans* with vaginal epithelial cells leading to DAMP release that can catalyse immunopathology during vaginal infection (Richardson *et al.*, 2018).

In contrast to the highly diverse, species-specific and damage-dependent transcriptional profiles to the four *Candida* species at late stages, the epithelial response was highly uniform at early stages. We identified that the initial uniform response of epithelial cells towards the different *Candida* species was driven by general epithelial-driven recognition mechanisms rather than directed by convergent activities (such as virulence programs) of the tested *Candida* species. For example, independent of their viability, expression of *DOCK8* was upregulated following infection by any of the four *Candida* species (supplementary files S5-S8, sheets “res_<species>_human_0_3 and res_<species>_human_0_3D”). Although this gene has multiple signalling functions, its role as an intermediate to promote immune responses to diverse external stimuli is emerging (Kearney, Randall and Oliaro, 2017). Several studies associated *DOCK8* with mucocutaneous candidiasis due to impaired Th17 differentiation (Choate, 2009; McGhee *et al.*, 2010; Chu, 2012). Thus, it is tempting to speculate that this gene is involved in regulating the recognition of *Candida* by epithelial cells.

A subset of uniformly up-regulated genes required both fungal viability and physical contact. Surprisingly, this included genes associated with the type I IFN response pathway. This pathway was generally thought to be associated with viral infections (‘Virus interference. I. The interferon’, 1957). Nevertheless, a previous study observed a type I IFN response by peripheral blood mononuclear cells infected with *C. albicans* (Smeekens *et* *al.*, 2013). Besides, type I IFN responses were recently shown to dysregulate host iron homeostasis and enhances *C. glabrata* infection (Riedelberger *et al.*, 2020) and the type I IFN-inducing RIGi Helicase MDA5 has been associated with systemic candidiasis as well as chronic mucocutaneous candidiasis (Jaeger *et al.*, 2015). Finally, IFNαR1 signalling is crucial in a murine model for efficient host defence against systemic candidiasis (del Fresno *et al.*, 2013).

Our data show, for the first time, that type I IFN signalling is induced by human vaginal epithelium in response to *Candida* species with two different effects: dampening of pro-inflammatory responses and epithelial protection. Such an immune response may be relevant in host niches colonized by commensal microbes. For example, intestinal epithelial cells maintain the intestinal homeostasis *via* expression of type I IFN and ISGs, which regulate the durability and specificity of immune responses and guide the immune system to differentiate between commensal and pathogenic microbiota (Sato *et al.*, 1998; Munakata *et al.*, 2008; Kotredes, Thomas and Gamero, 2017). Since *Candida* spp. are commensals of vaginal mucosa (Pekmezovic *et al.*, 2019), it might be that the epithelial type I IFN pathway serves as a threshold regulating neutrophil driven antifungal immunity. Our results support such a protective role for type I IFN responses in increasing epithelial resistance to damage and induction of potentially detrimental neutrophil driven responses. Blocking of type I IFN signalling by IFNAR neutralization on vaginal epithelial cells resulted in pro-inflammatory responses in neutrophils, possibly resembling immunopathology state observed *in vivo*. While down-regulating pro-inflammatory responses, we also observed induction of type I IFN signalling by exogenous IFNβ which can increase epithelial resistance to damaging activities of *Candida* spp. in our model, potentially by inducing epithelial antifungal host defence. Supporting this, Li *et al*. showed that administration of recombinant human IFNα-2b decreased the inflammation and vaginal epithelial damage in an *in vivo* rat VVC model (T. Li *et al.*, 2019). These combined effects, immunomodulation and epithelial antifungal activities, may be crucial to restrict *Candida* spp. to commensalism and avoid pro-inflammatory immune responses leading to immunopathology.

The second group of genes triggered in the early infection stage, depending on both fungal viability and contact, were encoded by mitochondrial DNA (mtDNA), in particular genes coding for the respiratory electron-transport chain potentially indicating altered mitochondrial function. We confirmed that *Candida* infection changes the mitochondrial membrane potential of epithelial cells. A transient hyperpolarization of mitochondrial membrane accompanied by mtROS production was observed in epithelial cells exposed to all four *Candida* species, but not in uninfected control cells. This could potentially indicate higher energy demand associated with mitochondrial hyperactivation in the presence of the fungal pathogens.

Apart from the well-established roles of mitochondria in cellular metabolism and energy production, recent findings support the view that mitochondria are central hubs in innate immunity (Mills, Kelly and O’Neill, 2017; Mohanty, Tiwari-Pandey and Pandey, 2019). Mitochondrial dysfunction, resulting in mtROS production and mtDNA release to the cytosol, can act as a DAMP and thus activate different signalling pathways (Seth *et al.*, 2005; West *et al.*, 2015; Mills, Kelly and O’Neill, 2017; West and Shadel, 2017; Grazioli and Pugin, 2018; Mohanty, Tiwari-Pandey and Pandey, 2019). Altered mitochondrial function at early stages of infection has been reported for several bacterial pathogens, including *Chlamydia trachomatis* (Kurihara *et al.*, 2019), *C. pneumonia* (Käding *et al.*, 2017), *Listeria monocytogenes (Stavru et al., 2011)*, and the parasite *Toxoplasma* gondii (Syn *et al.*, 2017).

Using electron microscopy, we observed morphological alterations of shape and integrity of mitochondria in *Candida*-infected vaginal epithelial cells. This observation might mechanistically be explained by hyperpolarization of the mitochondrial membrane potential and mtROS production due to increased activity of the electron-transport chain, which in turn lead to dysfunctions in selected mitochondria and initiation of their removal by mitophagy.

Intriguingly, similar phenomena were observed in host mitochondria of different cell types upon bacterial and viral infections (West *et al.*, 2011; Kim *et al.*, 2014; Plataki *et al.*, 2019; Ramond *et al.*, 2019). One of the reported down-stream consequences of mitochondrial signalling *via* DAMPs triggered by microbial infections is the induction of cytokine production (Brokatzky *et al.*, 2019), including type I IFN responses (Fang, Wei and Wei, 2016). Levels of the mitochondrial DAMP, mtDNA, were significantly elevated at 6 hpi in the cytosol of vaginal epithelial cells infected with all tested *Candida* species. To link the release of the cytosolic DAMP mtDNA with type I IFN responses, we demonstrated that transfection of uninfected cells with this mtDNA induced type I IFN responses. This provides evidence that mtDNA can act as a DAMP activating a type I IFN pathway in *Candida* infections of the vaginal epithelia. This activation may potentially occur through the stimulator of interferon genes (STING)-pathway, as shown for *Strepto*coccus pneumoniae (Gao *et al.*, 2019), on a level of post-translational modifications (P.-H. Wang *et al.*, 2018).

In order to maintain epithelial integrity and the capacity to mount epithelial host defence while preventing detrimental neutrophil-driven immune responses, such a mitochondrial response must occur on a sub-lethal level (Brokatzky *et al.*, 2019) and not lead to the activation of programmed cell death (apoptosis). Accordingly, we did not observe any sign of apoptosis neither during the early or the later stage of infection (Fig. 4.5, suppl. Fig. S4.4). The non-lethal mitochondrial dysfunction observed at early stages during infection with all *Candida* species was independent of measurable epithelial damage (since epithelial damage was only observed at later stages). Likewise, we saw consistent activation of the type I IFN pathway, which is normally suppressed by apoptotic caspases (Ning *et al.*, 2019). In fact, when we induced apoptotic signalling experimentally, induction of type I IFN was reduced. Similar infection studies with *L. monocytogenes* and *T. gondii* showed that mitochondrial dysfunction was uncoupled from the apoptotic pathway (Stavru *et al.*, 2011; Syn *et al.*, 2017). The mitochondrial apoptosis apparatus can be activated during infection of epithelial cells with diverse microbial agents at a low level, which is insufficient to induce apoptosis (Ichim *et al.*, 2015; Brokatzky *et al.*, 2019). This phenomenon has been termed limited mitochondrial outer membrane permeabilization, or “minority MOMP”, and induces pro-inflammatory cytokine production *via* STING.

Viruses, bacteria, and parasites are all able to cause minority MOMP, leading to DNA damage that depends on the mitochondrial apoptosis pathway, contributing to cytokine release during infection (Brokatzky *et al.*, 2019). We propose that this novel mechanism plays a significant role in epithelial sensing of *Candida* species, induction of epithelial antifungal immunity, and modulation of neutrophil-driven innate immune responses.

Collectively, we identified species-specific pathogenicity patterns of *Candida* species infecting vaginal epithelial cells, which are reflected on the transcriptional level during the course of infection. In contrast, vaginal epithelial cells exhibited a conserved response at early stages, but a diverse, damage-driven response at later stages. The conserved response was characterized by non-lethal mitochondrial signalling, which induced a type I IFN response, which protects epithelial cells against fungal-induced damage and causes modulation of innate antifungal immune responses. This acts as a novel common pathway of host-pathogen interactions between vaginal epithelial cells and *Candida* pathogens.

4.5 Materials and Methods

Fungal strains and culture conditions

C. albicans SC5314 (Gillum, Tsay and Kirsch, 1984), *Candida glabrata* ATCC® 2001™ (obtained from American Type Culture Collection; ATCC), *C. tropicalis* DSM 4959 (obtained from DSMZ-German Collection of Microorganisms and Cell Cultures), *C. parapsilosis* 73-037 (Tavanti *et al.*, 2005) and *C. albicans ece1*Δ/Δ (Moyes *et al.*, 2016) were used in this study. For all experiments, single colonies were picked from Yeast Peptone Dextrose (YPD) agar plates and grown overnight in liquid YPD medium in an orbital shaker at 180 rpm at 30 °C (*C. albicans, C. tropicalis, and C. parapsilosis*) or 37 °C (*C. glabrata*). Yeast cells were then harvested by centrifugation (20 000 g, 1 min), washed twice with phosphate-buffered saline (PBS), and adjusted to $2\text{x}10_6$ yeast cells per millilitre (yeast ml_{-1}).

In vitro vaginal epithelial infection model

To mimic the vaginal epithelium, A431 epithelial cells (ECs) (DSMZ no. ACC 91) were used. These cells are derived from a vulva epidermoid carcinoma and routinely used to model the vaginal mucosa (Hernandez and Rupp, 2009; Schaller and Weindl, 2009). ECs were cultivated in RPMI-1640 medium (ThermoFisher Scientific) supplemented with 10% fetal calf serum (FCS, Bio&Sell) in a humidified incubator at 37°C, 5% CO_2. For the infection, ECs were seeded in 6-well plates ($3\times10_5$ cells per well) and cultured for 2 days. On the day of infection, media in each well was replaced with 1.5 ml RPMI-1640 without FCS and incubated for 30 min in order to allow cells to adjust to the change of medium. In subsequent bioinformatics analyses, we considered the control samples at 30 min in this medium as the 0 h time point. ECs were subsequently infected with *Candida* cells (1.5 ml of $2\times10_6$ yeast ml_{-1} in RPMI-1640 without FCS) and incubated at 37°C, 5% CO_2. Samples for RNA isolation were collected at different time points: 1.5 (*C. albicans* only), 3, 12, and 24 ≥ post-infection (hpi). More specifically, the well content was removed and replaced with 500 µl of RNeasy Lysis (RLT) buffer (Qiagen), containing 1% β-mercaptoethanol (Roth). Cells were detached using a cell scraper (< 3 min), immediately shock-frozen in liquid nitrogen, and stored at -80°C until further use (see “RNA isolation”). As controls, *Candida* cells alone and ECs alone were incubated for 30 min (0 h control: C0) and 24 hours (24 h control: C24) and samples for RNA isolation were collected as described above.

RNA isolation

Collected samples were defrosted on ice and centrifuged for 10 min (20 000 g, 4°C). The supernatant was transferred to a new vial and used to isolate human RNA (RNeasy Mini Kit, Qiagen), according to the manufacturer's instructions. Fungal RNA was isolated from the pellet, using a freezing-thawing method, as described previously (Wächtler *et al.*, 2011). Both human and fungal RNA concentrations were quantified using a NanoDrop 1000 Spectrophotometer (ThermoFisher Scientific) and RNA quality was assessed with Agilent 2100 Bioanalyzer (Agilent Technologies). Fungal and human RNA samples were subsequently pooled in a 2:3 quantitative ratio by weight.

UV killing of Candida

Candida cells from overnight cultures were collected by centrifugation, washed twice with PBS and adjusted to approximately 5×10$_7$ yeast ml$_{-1}$ in PBS. The suspension was transferred to a Petri dish as a thin liquid layer (10 ml) and exposed to 4 doses of 100-120 mJ per cm$_2$ in a UV-crosslinker (CL 508 S, Uvitec, Cambridge). The efficiency of UV killing was evaluated by plating 50 µl of the sample onto YPD agar and incubated for 48 h at 30°C.

Adhesion assay

ECs were infected with *Candida* yeast cells as described above and incubated for 1 hour. Non-adherent *Candida* cells were removed by rinsing with PBS. Subsequently, ECs with adhered *Candida* were fixed with Roti®-Histofix 4 % (Roth). Adherent *Candida* cells were stained with Alexa Fluor 647 conjugate of succinylated concanavalin A (ConA; Invitrogen) and visualized using a fluorescence microscope (Leica DM5500B, Leica DFC360 FX). The number of adherent cells was determined by counting at least 100 high power fields and expressed as a percentage of adherent cells (Wächtler *et al.*, 2011).

Invasion assay and hyphal length

ECs were infected with *Candida* cells as described above and incubated for 3 h. Non-adherent *Candida* cells were removed by rinsing with PBS and ECs and invading *Candida* were fixed with Roti®-Histofix 4 %. Extracellular, non-invasive fungal components were stained by ConA. After rinsing with PBS, ECs were permeabilized in 0.5% Triton X-100 for 10 min. Next, fungal cells were stained with Calcofluor white (CFW; Sigma-Aldrich) and visualized by fluorescence microscopy. The total hyphal length was noted, as well as the percentage of invasive hyphae (only CFW-stained), counted from at least 100 hyphae per strain.

Epithelial damage assay

ECs were infected with *Candida* cells as described above and incubated for 24 h. Epithelial damage was quantified by measuring lactate dehydrogenase (LDH) release using a Cytotoxicity Detection Kit (Roche) according to the manufacturer's instructions. The background LDH value of uninfected ECs (low control) was subtracted, and the corrected LDH release was expressed as % of high (full lysis) control (maximum LDH release induced by the addition of 0.25% Triton X-100 to uninfected ECs) unless otherwise stated. For the protection effect experiments, 0.1 ng/ml of interferon-beta (IFNβ; Invivogen) was added to ECs one hour prior to infection.

Transwell assay

ECs were seeded in 24-well plates ($1\times10_5$cells per well) in RPMI-1640 with FCS and incubated for two days at 37°C and 5% CO_2. The medium was exchanged with 750 µl of RPMI-1640 without FCS. Transwell inserts (polycarbonate membrane inserts with 0.4µm pore size; Corning), loaded with 250 µl of *C. albicans* or *C. glabrata* suspension ($4\times10_6$ yeast ml_{-1}), were placed in the wells. After three hours of incubation, the inserts were discarded and human RNA samples were collected and isolated as described above.

Reverse transcription-quantitative PCR (RT-qPCR)

Isolated RNA (500 ng) was treated with DNase I (Fermentas) following the manufacturer's recommendations and subsequently transcribed into cDNA using 0.5 µg Oligo(dT)12-18 Primer, 200 U Superscript™ III Reverse Transcriptase and 40 U RNaseOUT™ Recombinant RNase Inhibitor (Thermo Fischer Scientific). Obtained cDNA was diluted 1:5 and used for qPCR with GoTaq® qPCR Master Mix (Promega) in a CFX96 thermocycler (Bio-Rad). The expression levels were normalized against beta-actin or 18s rRNA. All the primers used are listed in Table S1.

Measurement of EC mitochondrial DNA (mtDNA) release

The release of mtDNA in response to infection was measured using the protocol of Bronner and O'Riordan (Bronner and O'Riordan, 2016) with some modifications. Briefly, ECs were seeded in 6-well plates and infected as described above. After 6 h of infection, the well content was removed and 200 µl of 1% Igepal CA-630 (NP-40; Sigma-Aldrich) was added and cells were scraped. Lysates were incubated on ice for 15 min and centrifuged (12 000 g, 15 min, 4 °C). The supernatant was used to isolate human mtDNA from the cytosolic fraction using the DNeasy Blood & Tissue Kit (Qiagen), according to the manufacturer's instructions. Finally, quantitative PCR is employed to measure cytosolic human mtDNA using

18s rRNA as a reference (Bronner and O’Riordan, 2016). Obtained results for infected ECs were compared to the values of the uninfected control. Tunicamycin(10 µM; Sigma-Aldrich) was used as a positive control, as ER-stress inducer that leads to mitochondrial dysfunction (Win *et al.*, 2014). The same procedure was carried out on yeast cells only. The lysis step did not cause any lysis of yeast cells and no DNA was detected after the isolation procedure, indicating that the DNA obtained was originating from epithelial cells only.

Measurement of EC mitochondrial membrane potential (ΔΨm)

Mitochondrial membrane potential (ΔΨm) was assessed using a JC-1 Mitochondrial Membrane Potential Assay Kit (Abcam). Tetraethylbenzimidazolylcarbocyanine iodide (JC-1) is a cationic dye that yields green fluorescence (emission 530±15 nm) in monomer state and red to orange (emission 590±17.5 nm) in the aggregate state. Since aggregates form due to high mitochondrial membrane potential, this allows to measure changes towards depolarization or hyperpolarization in infected cells and compared to uninfected control. ECs were seeded and infected in 96-well black clear-bottom plates and the measurements were done 1, 3, 6, and 24 hpi, according to the manufacturer’s instructions. Carbonyl cyanide 3-chlorophenylhydrazone (CCCP), a protonophore that can cause mitochondrial depolarization, was used as a positive control (100 µM).

EC mitochondrial reactive oxygen species (mtROS) detection

Production of mtROS was measured using a Mitochondrial Reactive Oxygen Species Detection Assay Kit (Cayman Chemicals). ECs were seeded and infected in a black clear-bottom 96-well plate and measurements of infected and uninfected cells were done 1, 3 and 6 hpi, according to the manufacturer’s instructions. Antimycin A (100 µM), inducing superoxide radicals leakage from mitochondria, was used as a positive control.

Transfection of ECs with the cytosolic DNA

Cytosolic EC DNA (2 µg ml$_{-1}$) was isolated from infected and uninfected ECs as described above and used for transfection of fresh ECs using UltraCruz® Transfection Reagent (Santa Cruz) according to the manufacturer's instructions. After 6 and 24 h, RNA samples were collected and the expression of ISGs was quantified using qPCR. Transfection with cytosolic DNA from the uninfected control was used as a mock transfection.

Apoptosis/necrosis assay

ECs were seeded in 96-well plates and infected with *Candida* cells as described above. Simultaneously with infection, ECs were stained for apoptosis, by measuring caspase 3/7 activity (CellEvent™ Caspase 3/7 detection reagent; Invitrogen, Life Technologies) and necrosis, using propidium iodide (PI; Sigma), according to the manufacturer's instructions. ECs were imaged 3 and 24 hpi using a Cell Discoverer 7 (Carl Zeiss) with fluorescence settings at 488/520 nm (caspase 3/7) and 535/617 nm (PI). The fluorescence was measured according to the manufacturer's instructions to obtain quantitative data. Staurosporine (1.2µM; Sigma) was used as a positive control, while uninfected ECs were used as a negative control.

Apoptosis induction

ECs were seeded in 96-well plates and infected with *Candida* cells as described above, with the addition of staurosporine (1.2µM) simultaneously with infection. After 6h, RNA samples were collected and the expression of ISGs was quantified using qPCR. Results were compared to infected cells incubated in the media without staurosporine.

Collection of EC supernatants

ECs were infected with *Candida* cells as described above and incubated for 24 h, in the presence or absence of neutralizing anti-human IFNAR2 antibody (1 µg ml$_{-1}$, PBL Inferon Source, Piscataway, USA). Supernatants were collected and stored at -80 °C until use (see Cytokine release and Neutrophil stimulation). Supernatants of *Candida* cells only were included as control.

Cytokine release

ECs were infected with *Candida* cells as described above and incubated for 24 h. The release of interleukin(IL)-6 , IL-8, IL-1α and IL-1β was measured by commercially available human enzyme-linked immunosorbent assay (ELISA) kits (IL-6, IL-8, IL-1β: Invitrogen; IL-1α: R&DSystems) according to the manufacturer's instructions.

Blood donors

Human peripheral blood was collected from healthy volunteers with ethics approval and after obtaining written informed consent. This study was conducted according to the principles expressed in the Declaration of Helsinki. The blood donation protocol and use of blood for this study were approved by the institutional ethics committee of the University Hospital Jena (permission number 2207-01/08).

Neutrophil cytokine production

Primary human neutrophils were isolated from blood using a previously published protocol (Gresnigt *et al.*, 2012) and seeded in a 24-well plate ($5x10_5$ cells ml_{-1}). The supernatants from infected (24 hpi) and uninfected ECs were used to stimulate neutrophils for 24 h. After incubation cytokine release was measured using ELISA as described above.

TEM imaging

The cells were fixed by adding glutaraldehyde (2.5 % (v/v) final) to the growth medium. After 1 h the cell layer was gently scraped off the surface, collected as a pellet by centrifuging at 600 g, and washed 3 x with PBS. After post-fixation in osmiumtetroxide (1% (w/v) in aqua dest.) for 1 h, dehydration in ascending ethanol series with post-staining in uranylacetate was performed. Afterwards the samples were embedded in epoxy resin (Araldite) and ultrathin sectioned (60 nm) using an ultramicrotome Leica Ultracut E (Leica, Wetzlar, Germany). After mounting on filmed Cu grids and post-staining with lead citrate the sections were studied in a transmission electron microscope (EM 902A, Zeiss, Oberkochen, Germany) at 80 kV. Images were acquired with a 1k FastScan CCD camera (TVIPS, München).

RNA-Seq library preparation and sequencing

Library preparation for RNA-Seq was performed with the TruSeq Stranded mRNA Sample Prep Kit v2 (ref. RS-122-2101/2, Illumina) according to the manufacturer's instructions unless specified otherwise. One µg of total RNA was used for poly(A)-mRNA selection using streptavidin-coated magnetic beads. Samples were then fragmented to ~300bp and subsequently, cDNA was synthesized using reverse transcriptase (SuperScript II, Invitrogen) and random primers. The second strand of the cDNA incorporated dUTP in place of dTTP. Double-stranded DNA was further used for library preparation. It was subjected to A-tailing and ligation of the barcoded Truseq adapters. All purification steps were done using AMPure XP beads (Agencourt). Library amplification was performed by PCR on the size selected fragments using the primer cocktail supplied in the kit. Final libraries were analyzed using Agilent DNA 1000 chip (Agilent) to estimate the quantity and check fragment size distribution and were then quantified by qPCR using the KAPA Library Quantification Kit (KapaBiosystems) before amplification with Illumina's cBot. To avoid potential batch effects, all samples were randomly distributed on the sequencing flowcells. Libraries were sequenced with 2x50 (n=21), 2x75 (n=74) and 2x150 (n=1) read lengths on Illumina's HiSeq 2500 (2x50 bp) and HiSeq 3000 (the rest) at the Genomics Unit of the Centre for Genomic Regulation, Barcelona, Spain. Raw sequencing data have been deposited in SRA under the accession numbers SRR10279972-SRR10280067.

Bioinformatics data analysis

FastQC v0.11.6 *(*http://www.bioinformatics.babraham.ac.uk/projects/fastqc*)* and Multiqc v. 1.0 (Ewels *et al.*, 2016) were used to perform quality control of raw sequencing data. Read trimming, when necessary, was performed by Trimmomatic v. 0.36 (Bolger, Lohse and Usadel, 2014) with TruSeq3 adapters using 2:30:10 parameters and discarding reads shorter than sequenced read length. To visualize read alignments, we used the Integrative Genomic Viewer v. 2.3.97 (IGV) (Robinson *et al.*, 2011).

For read mapping and quantification, we used splice-junction sensitive read mapper STAR v 2.5.2b (Dobin *et al.*, 2013) using basic two-pass mode and default parameters. For samples comprising either fungal or human RNA, reads were mapped to the corresponding reference genomes. In the case of pooled samples containing RNA from both the host and the pathogen, the data were mapped to concatenated human and corresponding yeast reference genomes. For human data, we used the primary genome assembly GRCh38 and genome annotations from Ensembl database release 89 (last accessed on 8 of August 2017) (Hunt *et al.*, 2018). Reference genomes and genome annotations for *C. albicans* SC5314 (assembly 22), *C. glabrata* CBS138 and *C. parapsilosis* CDC317 were obtained from *Candida* Genome Database (CGD, last accessed on 17 of August 2017, (Skrzypek *et al.*, 2017). From the phased reference genome and annotations of *C. albicans*, we selected Haplotype A to perform further analysis in order to avoid substantial rates of ambiguously mapped reads. Reference sequence and annotations for *C. tropicalis* were obtained from RefSeq database (last accessed on 9 of August 2017, (O'Leary *et al.*, 2016). The genes missing from RefSeq genome annotations were manually added from CGOB database (Maguire *et al.*, 2013). GFF genome annotation files were converted to GTF format using gff read utility v. 0.9.8 (Trapnell *et al.*, 2010).

We used Centrifuge v. 1.0.4 (Kim *et al.*, 2016) to test the presence of contamination in our dataset, by remapping the reads that did map neither to human nor fungal reference genomes to the whole NCBI nt database (downloaded on the 23rd of March 2018).

To assess the rates of cross-mapping, i.e. reads originated from human but mapped to fungi and vice-versa, which can bias expression levels quantifications, we used Crossmapper v. 1.1.0 (Hovhannisyan, Hafez, *et al.*, 2020) which simulates reads from multiple reference genomes/transcriptomes, maps the data back to the concatenated reference sequences and reports the rates of cross-mapping. We used the "RNA" mode of Crossmapper and simulated and back-mapped 20 and 40 million 2x50 and 2x75 reads for each fungal species and human, respectively.

Differential gene expression analysis was performed using the Bioconductor package DESeq2 v. 1.22.2 (Love, Huber and Anders, 2014) using read counts obtained by STAR mapping. For human samples and each fungal species, we compared time point 0 with other time points throughout the course of infection by Wald test using *contrast* option of DESeq2. To detect any statistically significant changes of expression throughout the course of infection, we also used a likelihood ratio test of DESeq2, by dropping the "time" component of the formula design. Genes with |log2 fold change|>1.5 and adjusted p-value (p_{adj}) < 0.01 were considered differentially expressed, unless specified otherwise. To account for possible batch effects in the experiments involving *C. albicans ece1Δ/Δ* and non-viable fungal cells, we applied RUVg function of RUVseq v1.16.1 (Risso *et al.*, 2014) Bioconductor package, using non-differentially expressed genes (basemean>10 and p_{adj}>0.05) across all samples and time points as negative controls. Since the optimal parameters for the batch effect removal algorithm are not defined in prior, we employed a strategy of incremental increase of k values (k=1,2,...,n), until we observed disruption of the PCA clustering of original data from the first batch of sequencing. To perform differential expression analysis, the obtained matrix of batch effect coefficients was further supplied to the design formula of DESeq2 object, which was subsequently run using original count data. For plotting “batch-free” PCA plots, we used batch-corrected counts retrieved from RUV package.

The list of 1-to-1 orthologs between the four fungal species was obtained from CGD. For interspecies gene expression comparisons, the raw read counts for each fungal species were normalized by library size and gene length. Gene Ontology (GO) term enrichment analysis was performed using clusterProfiler package v. 3.10.1 (Yu *et al.*, 2012). GO information for fungal species was obtained from CGD, while for human data we used "Genome-wide annotation for Human database” (org.Hs.eg.db) v. 3.7.0 in R (http://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html). All down-stream data analysis and visualization were performed in R v. 3.5.3 using various packages.

Statistical analysis

Experiments were performed at least in biological triplicates (n≥3) with at least three different donors (neutrophil cytokine release) or three independent experiments. TEM and fluorescence microscopy were done only once. Data were analysed using GraphPad Prism 8 (GraphPad Software, La Jolla California USA). Values are presented as mean ± standard deviation (SD). All the ratio data were ln transformed as indicated prior to statistical analysis in GraphPad Prism and compared to 0 (uninfected control) using one-sample t-test, except for Fig. 4.6C where unpaired t-test was used. Statistically significant results are indicated by asterisks as follows: *, $p \leq 0.05$; **, $p \leq 0.01$.

Data availability

Raw sequencing data have been deposited in SRA under the accession numbers SRR10279972-SRR10280067.

Code availability

All custom scripts are available at our GitHub page https://github.com/Gabaldonlab/Host-pathogen_interactions for result reproducibility.

Acknowledgements

M.P., H.H., E.I., J.O.P., T.G., G.B. and B.H. received funding from the European Union Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 642095 (OPATHY). B. H. further receives support from the European Union Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 812969 (FunHoMic) and the Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Centre (CRC)/Transregio 124 FungiNet (Project C1). M.S.G. was supported by a Humboldt Research Fellowship (Alexander von Humboldt-Stiftung). We would like to thank Christine Kämnitz from Electron Microscopy Center in Jena for sample preparation in TEM.

4.6 Supplementary figures

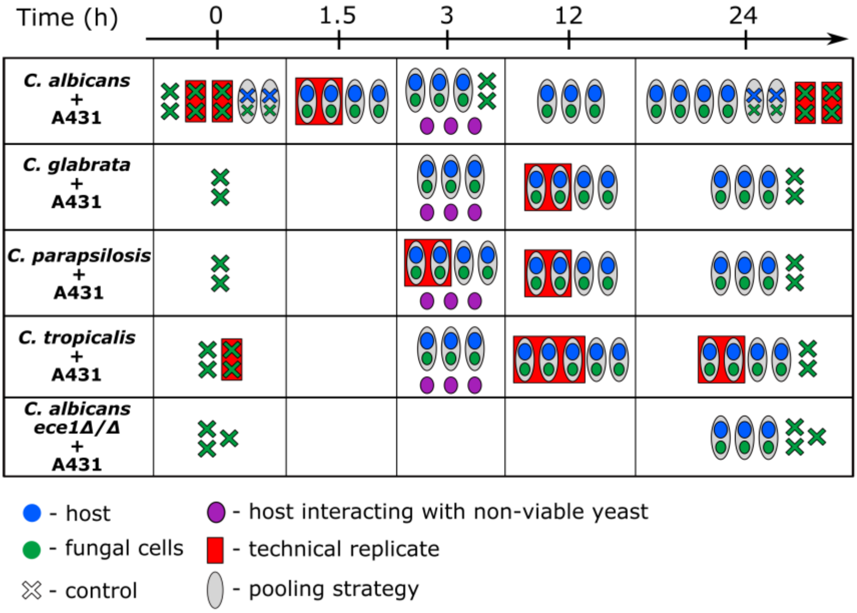


Fig. S4.1. Overall experimental design of the current study. Each symbol corresponds to a sequenced sample. Host samples are depicted in blue; *Candida* samples are depicted in green; control samples (human or fungal cells only) are indicated by crosses; human samples interacting with non-viable fungal cells are shown with purple ovals; technical replicates (i.e. the same sequencing library sequenced several times) are highlighted in red, and the strategy for combining human and fungal RNA is shown with grey ovals. The arrow at the top indicates the time-course of the infection in hours (h).

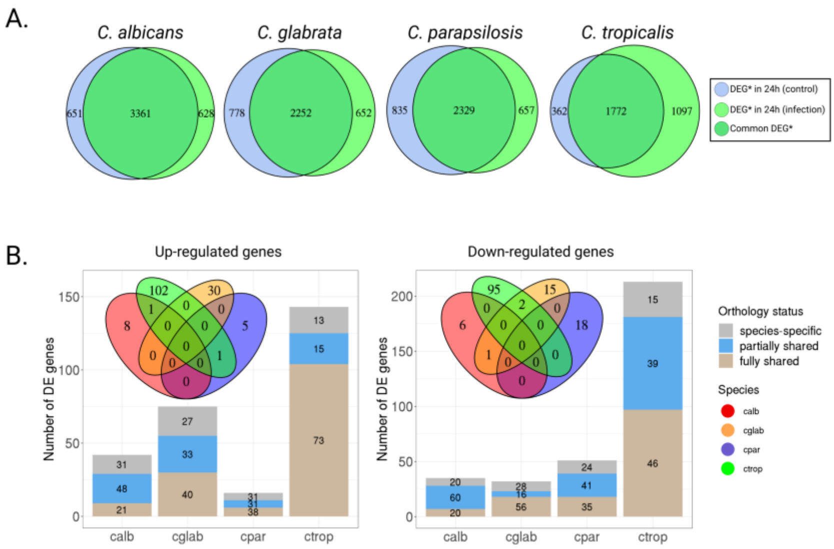


Fig. S4.2. Differentially expressed (DE) genes of yeast species at 0 vs 24 hpi and 0 vs 24-hour control (hc). For detecting infection-specific genes with a higher stringency, we used filters of log2fold change>0 and padj<0.01. Then for identified genes, we applied the |log2fold change| > 1.5 for downstream analysis to be consistent with other results. **(A)** Venn diagrams indicating similarities and differences of fungal DE genes in culture medium only (control) and in response to epithelial cells (infection). **(B)** Distribution of infection-specific fungal genes across studied Candida pathogens. Bar plots demonstrate the distribution of partially shared, fully shared and species-specific genes. Numbers on bar plots indicate the percentage (%). Venn diagrams depict numbers of fully shared genes across species. “calb” denotes C. albicans, “cglab” - C. glabrata, “cpar” - C. parapsilosis and “ctrop” - C. tropicalis.

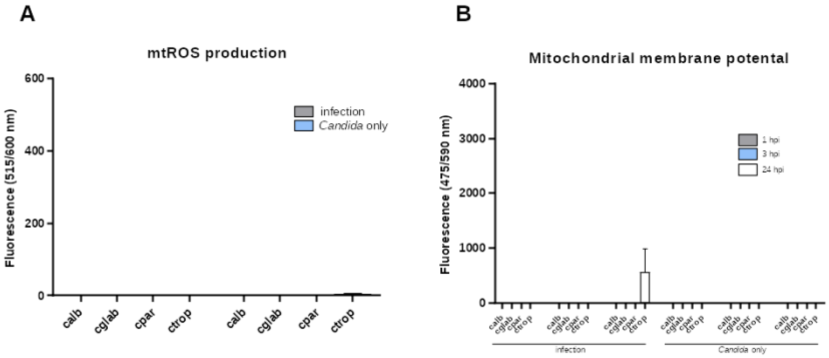


Fig. S4.3. Fungal-only controls. (A) mtROS production and (B) Mitochondrial membrane potential measured during the infection and in fungal-only cultures. All values are presented as mean ± SD, n=3. Labels of the data points correspond to

sample IDs; “calb” denotes *C. albicans*, “cglab” - *C. glabrata*, “cpar” - *C. parapsilosis*, “ctrop” - *C. tropicalis.*

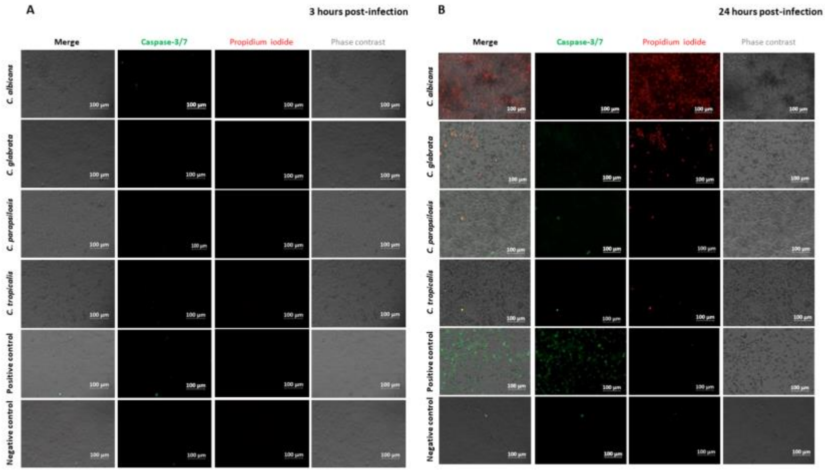


Fig. S4.4. Apoptosis and Necrosis of vaginal epithelial cells infected with *Candida* species. Visualization of apoptosis (caspase 3/7; green) and necrosis (propidium iodide; red) in ECs infected with all four *Candida* species at **(A)** 3 hours post-infection and **(B)** 24 hours post-infection. (positive control: apoptosis inducer staurosporine (1.2 µM); negative control: uninfected ECs); n=1.

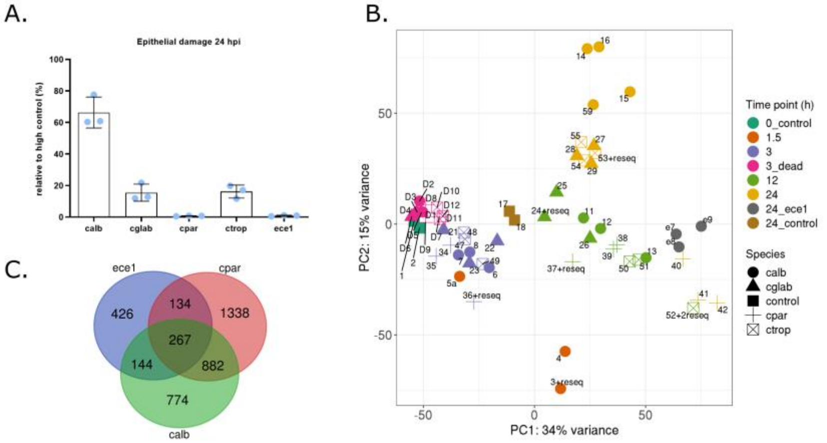


Fig. S4.5. Fungal damage-driven response of human transcriptome profiles. **(A)** Levels of LDH release by EC cells upon the damage by four fungal pathogens 24 hpi. All values are presented as mean ± SD, n≥3. **(B)** PCA plot of human samples interacting with non-viable and viable fungal species, including *C. albicans ece1Δ/Δ*. The plot is obtained using RUVg k=1 (see Fig. S4.6 for plots with alternative k values) **(C)** Venn diagram indicating similarities and differences of human DE genes in response to *C. albicans, C. albicans ece1Δ/Δ* and *C.*

parapsilosis. “calb” denotes human samples interacting with *C. albicans*, “cglab” - with *C. glabrata*, “cpar” - with *C. parapsilosis*, “ctrop” - with *C. tropicalis,* and “ece1” - *C. albicans ece1Δ/Δ*.

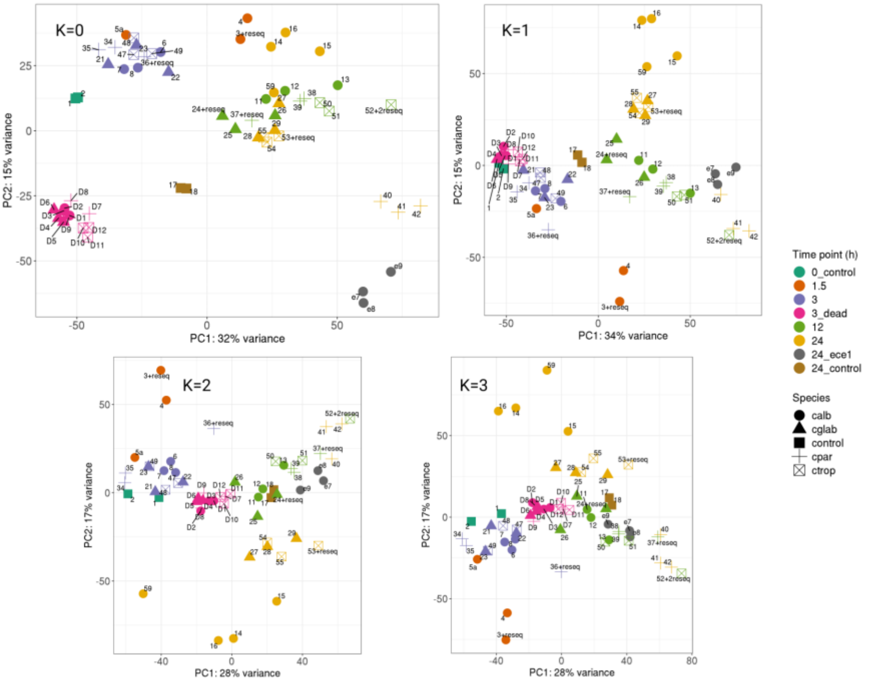


Fig. S4.6. Fungal damage-driven response of human transcriptome profiles. PCA plot of human samples interacting with fungal cells obtained using k=0,1,2,3 of RUVseq package for batch effect correction. “calb” denotes human samples interacting with *C. albicans*, “cglab” - with *C. glabrata*, “cpar” - with *C. parapsilosis*, “ctrop” - with *C. tropicalis*, and “ece1” - *C. albicans ece1Δ/Δ.*

5 The long non-coding RNA landscape of *Candida* pathogens

5.1 Abstract

Long non-coding RNAs (lncRNAs) is a poorly studied class of transcripts with emerging important roles in key cellular processes. Despite efforts to characterize lncRNAs across a wide range of species, these molecules have been poorly explored in eukaryotic microbes, including yeast pathogens of the *Candida* clade. Here, by mining thousands of publicly available RNA sequencing datasets, we inferred and analyzed comprehensive lncRNA repertoires of the four most common *Candida* pathogens: *C. albicans, C. glabrata, C. parapsilosis,* and *C. tropicalis*. We show that each of these species have hundreds of lncRNAs, which compared to protein coding genes are shorter, have lower GC content, and lower expression levels. The lncRNAs showed low primary sequence conservation across the studied species, but shared significant levels of synteny and secondary structure conservation. Our analyses revealed widespread patterns of co-expression between coding and non-coding transcripts. Finally, we traced the expression of these transcripts through the course of human epithelial infection, and identified lncRNAs specifically expressed during infection. Our work presents the first comprehensive assessment of lncRNA landscapes in *Candida*, and paves the way for research of the role of lncRNAs in the major human yeast pathogens.

Key words: lncRNAs, *Candida* yeasts, host-pathogen interactions

5.2 Introduction

Advances in high-throughput RNA sequencing in the past decade have shown that eukaryotic cells express abundant and numerous types of non-coding transcripts (Nagalakshmi *et al.*, 2008; Hangauer, Vaughn and McManus, 2013; Kung, Colognori and Lee, 2013; Zhao *et al.*, 2016; Till, Mach and Mach-Aigner, 2018; Uszczynska-Ratajczak *et al.*, 2018). One type of non-coding transcripts are long non-coding RNAs (lncRNAs), broadly defined as transcripts longer than 200 bp which do not code for proteins. These molecules have several peculiarities, which are consistently reported across a wide range of taxa - they are expressed at low levels and with a high cell type specificity (Cabili *et al.*, 2011). As compared to protein-coding genes, lncRNAs are poorly conserved at the sequence level, show a rapid evolutionary turnover, and are often species-specific (Kutter *et al.*, 2012; Paralkar *et al.*, 2014; Sarropoulos *et al.*, 2019). It has been suggested that functionality of lncRNAs can be attributed to their secondary structure (Johnsson *et al.*, 2014; Guo *et al.*, 2016; Zampetaki, Albrecht and Steinhofel, 2018), which can be maintained by selective pressures (Pegueroles and Gabaldón, 2016). However, whether lncRNAs are highly structured is still debated (Spitale *et al.*, 2015; Yang and Zhang, 2015). LncRNAs have been shown to play important roles in numerous processes such as, among others, gene expression regulation, imprinting, splicing, and cell cycle (Merry, Niland and Khalil, 2015; Jandura and Krause, 2017; Zhang *et al.*, 2019). However, despite extensive research, only a limited number of lncRNAs has been functionally characterized.

Non-coding transcripts are also abundant in fungi and have been shown to regulate various processes including cell wall remodeling, transcriptional control, and response to nutrients (David *et al.*, 2006; Nagalakshmi *et al.*, 2008; Donaldson *et al.*, 2017; Novačić *et al.*, 2020). Most of our knowledge regarding fungal lncRNAs comes from the model organism *S. cerevisiae* (Niederer, Hass and Zappulla, 2017; Till, Mach and Mach-Aigner, 2018; Novačić et al., 2020), but research on lncRNAs has recently expanded to other fungi, such as *Neurospora crassa* (Cemel *et al.*, 2017), *Fusarium graminearum* (Kim *et al.*, 2018), *Metarhizium robertsii* (Z. Wang *et al.*, 2019), *Pichia pastoris* (Sun *et al.*, 2019), the white-rot fungus *Ganoderma lucidum* (J. Li *et al.*, 2014) and the brown-rot fungi *Coniophora puteana* and *Serpula lacrymans* (Borgognone *et al.*, 2019). In accordance with findings for other taxa (Lopez-Ezquerra, Harrison and Bornberg-Bauer, 2017; Pegueroles, Iraola-Guzmán, *et al.*, 2019), these studies on diverse fungi consistently show that lncRNAs are generally shorter, have lower expression levels and GC content as compared to protein-coding genes. Some of these studies suggested novel putative functional implications of lncRNAs. For example, in the filamentous fungus *M. robetsii*, Wang et al., (2019) identified 1081 lncRNA transcripts that were differentially expressed due to heat shock, hinting to their potential implications in thermal stress response. In *F. graminearum* the expression of numerous lncRNAs was shown to be regulated in a stage-specific manner during the fruiting body development (Kim *et al.*, 2018).

A very limited number of studies, however, has been performed to characterize the role of lncRNAs in fungal virulence. Wang and colleagues (Y. Wang *et al.*, 2019) addressed this issue in the insect pathogen *Cordyceps militaris*, where >4000 lncRNAs were shown to be dynamically expressed during the fungal development. Moreover, when the *xrn1*, the final gene of the nonsense-mediated decay pathway, determining the fate of lncRNAs, was knocked-out, the attenuation of virulence and growth rates was observed.

The role of fungal lncRNAs in pathogenicity towards the human host has been studied in the basidiomycete *Cryptococcus neoformans*. The lncRNA *RZE1* was shown to control yeast-hyphae morphological transition through regulating the expression and export of the *ZNF2* transcript, which encodes the key morphogenesis transcription factor (Chacko *et al.*, 2015). This was the first study providing evidence of the involvement of lncRNAs in fungal pathogenicity. However, the possible implications of lncRNAs in fungal virulence in other major fungal pathogens remains unknown.

Yeasts from the *Candida* clade are among the most widespread human fungal opportunistic pathogens. Up to 30 distinct, phylogenetically diverse, *Candida* species have been reported to infect humans, mainly when immunocompromised (Papon *et al.*, 2013; Gabaldón, Naranjo-Ortíz and Marcet-Houben, 2016). *Candida* infections represent a high burden for global healthcare. They range from common superficial infections, such as vaginal candidiasis affecting 75% of the female population (Sobel, 2007), to life-threatening invasive infection, with mortality rates reaching 70% (Flevari *et al.*, 2013; Klingspor *et al.*, 2015). Most *Candida* infections are caused by four species, namely *Candida albicans*, *Candida glabrata*, *Candida parapsilosis* and *Candida tropicalis,* which together account for 85-90% of the cases (Guinea, 2014). In recent years, numerous studies have been performed to investigate interactions between *Candida* pathogens and different hosts, at the level of gene expression by using various techniques, such as microarrays (Andes *et al.*, 2005; Walker *et al.*, 2009; Hebecker *et al.*, 2016) and transcriptome sequencing (Hovhannisyan and Gabaldón, 2019). However, for the fungal side, all these studies are mainly focused on protein coding genes, and to date, lncRNAs in *Candida* pathogens have never been systematically studied, and thus, their potential roles in fungal virulence are unknown.

Here, we applied large-scale comparative transcriptomics to identify lncRNAs in the four major *Candida* pathogens, using a vast dataset of more than 2500 RNA-Seq samples. We characterized the major properties of the identified lncRNAs, and assessed their evolutionary relationships. Finally, we investigated the expression of these transcripts throughout the course of epithelial infection, revealing transcripts potentially involved in pathogenicity.

5.3 Materials and Methods

Datasets

We used two datasets to define the lncRNA landscapes of the studied yeasts: one comprising all RNA-Seq data publicly available at the Short Read Archive (SRA) database (Leinonen *et al.*, 2011), and another one comprising a single large-scale RNA-Seq experiment including the four species (Chapter 4). We will refer to these datasets as B (Broad) and S (Specific) datasets, respectively. Both datasets include data for the four *Candida* species - *C. albicans* (calb), *C. parapsilosis* (cpar), *C. tropicalis* (ctrop), and *C. glabrata* (cglab).

The B dataset data was retrieved from the Sequence Read Archive (SRA) database (last accessed on 19th of July 2019, (Leinonen *et al.*, 2011)) using sratoolkit v. 2.9.6-1 with prefetch and fastq-dump functions. In total 2475 libraries were downloaded, of which 2177 for *C. albicans*, 129 for *C. parapsilosis*, 123 for *C. glabrata*, and 46 for *C. tropicalis*. FastQC v0.11.6 (Andrews S. 2010) and Multiqc v. 1.0 (Ewels *et al.*, 2016) were used to perform quality control of raw sequencing data. For *C. albicans*, we discarded 64 libraries with read length shorter than 49 bp. The remaining samples were pre-processed to obtain high-quality data. First, we trimmed all samples using Trimmomatic v. 0.36 (Bolger, Lohse and Usadel, 2014) with the following parameters: <ADAPTERS.fa>:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:49. Then, we assessed the strand-specificity of the libraries by first running RSEM prepare-reference v 1.3 to extract the transcriptomes of the analyzed species and then running salmon v. 0.8.1, which identifies the strandedness of the data (Patro *et al.*, 2017). We then discarded non-strand-specific libraries. This step ensures that expression of lncRNAs is not confounded by reads corresponding to other features (protein coding genes, tRNAs, rRNAs, etc.) located on the opposite DNA strand. With the remaining data, i.e. 666 libraries for *C. albicans*, 37 for *C. glabrata*, 71 for *C. parapsilosis* and 35 for *C. tropicalis*, we performed read mapping to the corresponding reference genomes using TopHat2 v. 2.1.1 with *--b2-very-sensitive* option (Kim *et al.*, 2013). Reference genomes and annotations for *C. albicans* SC5314 (assembly 22), *C. glabrata* CBS138 and *C. parapsilosis* CDC317 and *C. tropicalis* MYA-3404 were obtained from *Candida* Genome Database (CGD, last accessed on 17 of August 2017 (Skrzypek *et al.*, 2017)). Considering that the genome sequence of *C. albicans* is phased, in our analysis we used only haplotype A for read mapping to avoid a substantial amount of multi-mapped reads.

The S dataset corresponds to a previous RNA-Seq study of the interaction between human vaginal epithelial cell line A451 and the four *Candida* species (see Chapter 4). The dataset comprises samples taken at 1.5, 3, 12 and 24h post-infection, and includes controls for the effect of the culture medium on the fungal transcriptional activity. In total, the S dataset comprised 37 libraries for *C. albicans*, 14 for *C. glabrata*, 15 for *C.*

parapsilosis and 18 for *C. tropicalis*, representing strand-specific sequencing libraries with 2x50 and 2x75 bp read length. The samples which had traces of adapter sequences and/or poor quality bases were trimmed with Trimmomatic using <ADAPTERS.fa>:2:30:10 LEADING:1 TRAILING:1 SLIDINGWINDOW:4:1 MINLEN:<50,75> command. Subsequently, since most of the S dataset libraries comprised dual RNA-Seq data (i.e. mixed human and fungal RNA), we mapped the data to the concatenated reference genome of each fungus and human. Human reference genome GRCh38 and annotations were obtained from Ensembl database release 89 (last accessed on 8 of August 2017, (Zerbino *et al.*, 2018)). Then, we generated fungus-only bam files by subsetting mapped fungal reads from the pooled bam files using samtools v. 1.3.1 (Li *et al.*, 2009).

Computational prediction of lncRNAs

For each sample we performed genome-guided transcriptome assembly using Stringtie v. 1.3.3b (Pertea *et al.*, 2015). Then, for each species we merged all gtf files from all samples using stringtie merge with *-g 50* option which resulted in a unified transcriptome. This transcriptome annotation was compared with the original genome annotations of each species using gffcompare v. 0.11.2 to identify novel transcripts.

For *C. albicans*, our initial mapping and assembly strategy resulted in transcripts with artefactually long introns spanning several hundred thousand bases (data not shown). To avoid this, we repeated the analysis for all species setting the TopHat2 option --max-intron-length to 1000, which corresponds to approximate maximal intron length in fungi (Kupfer *et al.*, 2004). Additionally, for *C. albicans* we removed 4 assemblies from the project PRJNA292429 (assembled from SRR2153488-SRR2153491 accession numbers), which were producing very long intergenic transcripts compared to the other datasets thereby resulting in the bridging of novel intergenic transcripts with coding genes during the transcriptome merging step.

Next, we selected novel intergenic (“i”) and antisense (“a”) transcripts (corresponding to “u” and “x” class-codes of gffcompare output, respectively) longer than 200 bp. When several isoforms were present, we kept only the longest one. Further, we assessed the coding potential of the predicted transcripts using the CPC v. 0.9 (Kong *et al.*, 2007) and Feelnc v. 0.1.1 (Wucher *et al.*, 2017) software. CPC was run against the UniProt database (https://www.uniprot.org/downloads, last accessed on 9 July 2019), and the output transcripts assigned with “noncoding” label were retained. For Feelnc, we used the sequences of protein coding genes with *shuffle* mode as training datasets for its Random Forest machine learning algorithm. Additionally, for *C. albicans* and *C. tropicalis* we removed transcripts (n=438 and n=4, respectively) containing ambiguous nucleotides because they produced errors in the software runs. The coding potential cut-offs were defined using a 10-fold cross-validation, as implemented in Feelnc. Transcripts identified as non-coding by both software tools were considered as lncRNAs. In addition, the transcripts discarded from Feelnc runs due to the presence of ambiguous nucleotides, but identified as non-coding by CPC were included in our final lncRNA datasets.

Overall expression levels of lncRNAs

To assess the overall expression levels of lncRNAs and compare them to those of protein coding genes, we calculated the read counts of both transcript categories in all analysed samples using Featurecounts v. 1.6.4. The count data was normalized by transcript length and library size, resulting in transcripts per million (TPM) values. Statistical significance of differences between mean expression levels of protein-coding genes and lncRNAs was tested using the Wilcoxon test.

lncRNAs gene family classification

To assess evolutionary relationships between predicted lncRNAs across species, we used several independent strategies, namely BLAST reciprocal hits, secondary structure similarity, and analyses of synteny. Considering that overlapping features of antisense lncRNAs can potentially influence the results, only intergenic transcripts were used for these analyses.

To define best reciprocal hits between all possible pairs of species we used BLASTn v.2.9. To this end, we built a custom BLAST database for the set of lncRNAs of each species. Then, each set of lncRNAs was aligned against each database with BLASTn using e-value cut-off of 1e-3, and -max_hsps 1 and -max_target_seqs 1, which selects only the best alignment between matched query-sequence pair. Best reciprocal hits were selected.

To assess the relatedness of the lncRNAs based on their secondary structures, we first used RNAfold v.2.4.14 from the ViennaRNA package (Lorenz *et al.*, 2011) to obtain secondary structures for all studied intergenic lncRNAs. Then, using Beagle v.0.2 (Mattei *et al.*, 2015) with the local alignment mode, the lncRNAs structures of each species were aligned against those of other species in a pairwise manner. Similarly to BLASTn best reciprocal hits approach, for each lncRNA we then selected the hit with maximal zScore (at least zScore>3) and p<0.01.

We classified intergenic lncRNAs into syntenic transcripts using a methodology developed and validated in Pegueroles *et al.* (2019). Briefly, we first obtained the information of 1-to-1 orthologs between protein-coding genes in the four *Candida* species from CGD. Then, we defined pairwise syntenic relationships between the lncRNAs of studied species using the *synteny_nematodesv4GH.py* (https://github.com/Gabaldonlab/Synthenic-Families, with minor modifications directed to correctly match the species names and transcript identifiers in our study. The script was run with parameters *3 3 1*, i.e. considering three protein coding genes at each side of a given lncRNA, a minimum overlap of three shared genes and a minimum of one common gene at each side of a lncRNA. The analysis identified syntenic lncRNAs between each pair of species. Finally, to cluster best reciprocal hits, results of secondary structure alignments and pairwise syntenic lncRNAs into lncRNA families across species, we used *classifyFamiliesv5_VennGH.py* script (https://github.com/Gabaldonlab/Synthenic-Families) from Pegueroles *et al.*, 2019.

Co-expression analysis

We assessed the patterns of gene co-expression across all intergenic lncRNA transcripts and protein coding genes using the weighted correlation network analysis approach implemented in WGCNA v. 1.68 (Langfelder and Horvath, 2008). As for gene family classification, antisense transcripts were discarded from this analysis. As recommended by the package developers, we used log2(TPM+1) as expression values. For each species, we first selected the β power values using the “picksoftThreshold” function (see suppl. Fig. S5.1) implying an unsigned network. The minimum β value reaching 80% of scale-free network topology was used for downstream analysis. To reach the optimal values of β, we removed samples comprising outliers as identified by inspection of principal component analysis (PCA) plots based on expression values (suppl. Fig. S5.2). Namely, we removed the following samples: SRP099169 and SRP083839 for *C. tropicalis*, SRP151798 and SRP041812 for *C. parapsilosis* and SRP065276 for *C. glabrata*. Additionally, we removed all genes that had TPM values < 0.1 in more than 80% of the remaining samples. We inferred modules in the WGCNA networks using 1-Topology Overlap Matrix values, and identified eigengenes (i.e. the first principal component of each module). Finally, we assessed network and module connectivities and identified hubs, as implemented in WGCNA with defaults parameters. For each identified module, we performed GO term enrichment analysis of protein coding genes using ClusterProfiler v.3.1.1. GO term association tables were obtained from CGD. All custom calculations and data visualizations were performed in R v. 3.5.1 using various packages and using Inkscape graphics editor.

LncRNA expression during epithelial cell infection

We assessed the expression of all predicted lncRNAs during the infection model of vaginal epithelial cells interacting with *Candida* species (see Chapter 4). For this, we first calculated the mapped read counts for lncRNAs and protein coding genes using Featurecounts v. 1.6.4 (Liao, Smyth and Shi, 2014). For each species we performed differential expression analysis using DESeq2 v 2_1.22.2 Bioconductor package (Love, Huber and Anders, 2014), by comparing each time point of infection with 0 hour time point control samples using the Wald test. Differential expression calls were performed with the count data of both protein coding genes and lncRNAs. lncRNAs with |log2 fold change (L2FC)| > 1.5, and padj (adjusted p-value) < 0.01 were considered as differentially expressed (unless specified otherwise).

5.4 Results and Discussion

LncRNAs cataloging and overall characterization

To define and characterize lncRNA catalogs of the four main *Candida* pathogens, we used a vast dataset of 2571 samples comprising all publicly available RNA-Seq sequencing libraries for these species (Broad dataset, B), and additional data from a large-scale *in vitro* host-pathogen interaction study of these species with human vaginal epithelial cells (Specific dataset, S, see Materials and Methods for details). For each individual species, genome-guided transcriptome assemblies were performed independently for B and S datasets. These were subsequently merged and analyzed to produce a final predicted catalogues of lncRNAs for each species (Fig. 5.1, see Materials and Methods for details).

We divided lncRNAs into intergenic (“i”), i.e. transcripts that do not overlap with protein coding genes or other features, and antisense (“a”), i.e. lncRNAs overlapping coding genes or other features on the opposite DNA strand. The overall number of lncRNAs for each species is depicted on Fig. 5.2A. Full lncRNAs catalogs, including fasta and bed files for all species are available in Supplementary Materials and are deposited in Candidamine.

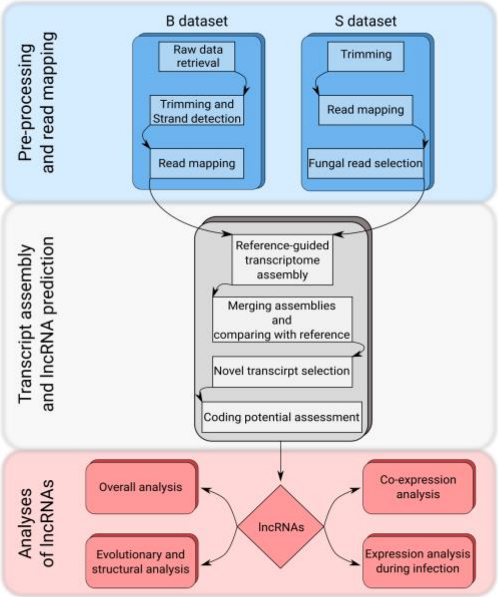

Pre-processing and read mapping
B dataset
Raw data retrieval
Trimming and Strand detection
Read mapping
S dataset
Trimming
Read mapping
Fungal read selection
Transcript assembly and lncRNA prediction
Reference-guided transcriptome assembly
Merging assemblies and comparing with reference
Novel transcirpt selection
Coding potential assessment
Analyses of lncRNAs
Overall analysis
Evolutionary and structural analysis
lncRNAs
Co-expression analysis
Expression analysis during infection

Fig. 5.1. Schematic representation of the bioinformatics workflow for lncRNA prediction and analysis. See description in the text.

The largest number of lncRNAs was detected in *C. albicans*, compising 5759 antisense and 1457 intergenic transcripts, followed by *C. parapsilosis* (3036 and 1498, respectively), *C. tropicalis* (987 and 1581, respectively) and *C. glabrata* (959 and 444, respectively). For species of the CTG clade (i.e. all except *C. glabrata*), the differences in the number of lncRNAs are mainly driven by the large number of antisense transcripts in *C. albicans*, which is about 2 and 6 times larger than observed in closely related *C. parapsilosis* and *C. tropicalis*, respectively. These differences may relate to the number of analyzed samples of these species - 699, 86 and 53 samples for *C. albicans*, *C. parapsilosis* and *C. tropicalis*, respectively.

To test this, we repeated the lncRNA prediction from subsets of different sizes, and produced saturation plots showing the dependency of the number of analyzed samples and the number of predicted lncRNAs (suppl. Fig. S5.3). The results of this analysis show that the number of antisense lncRNA in *C. albicans* reaches a plateau for subsets of ~200 samples and more. This suggests that few novel antisense lncRNAs in *C. albicans* may remain to be discovered. For the other species, considering they have at most 86 samples, it is likely that their antisense lncRNA catalogues are not complete, and might be expanded when new datasets are available.

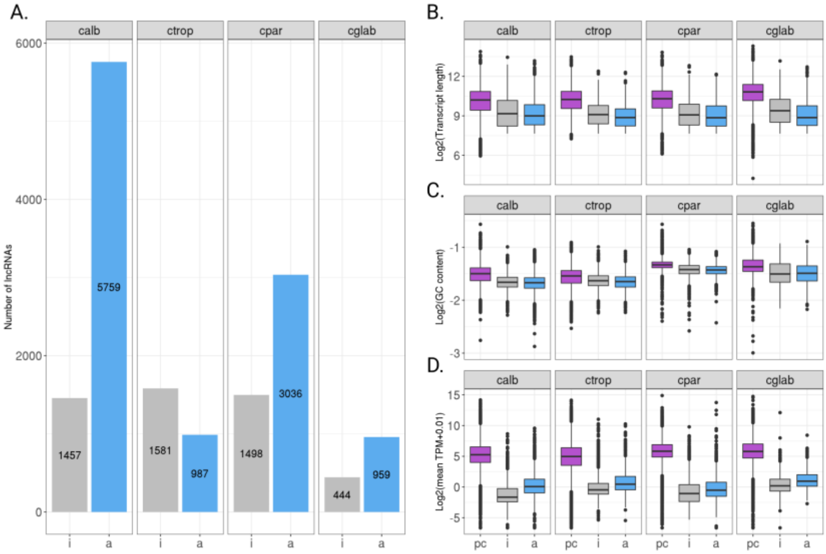


Fig. 5.2. The lncRNA landscapes and their molecular characteristics in four *Candida* species. **(A)** Overall distribution of intergenic (i) and antisense (a) lncRNAs in the studied pathogens; **(B)** Comparison of transcript lengths between lncRNAs and protein coding genes ("pc") across species; **(C)** Distribution of GC content of lncRNAs and protein coding genes across species; **(D)** Distribution of mean TPM values of lncRNAs and protein coding genes across species. For **(B)**, **(C)** and **(D),** all differences between lncRNAs and protein coding genes are statistically significant (Wilcoxon rank sum test p-value < 0.05).

In fact, *C. albicans* dataset comprises 125 different SRA project accession numbers, corresponding to different experiments, as compared to 10 and 12 accessions in case of *C. parapsilosis* and *C. tropicalis*. However, additional biological and technical factors that might influence this result, include pervasive transcription, transcriptional noise and varying efficiency of strand-specific library preparation protocols.

Of note, such drastic differences are not observed when analyzing only the samples of the host-pathogen interaction study (S dataset, 1088, 627, 968 antisense lncRNAs in *C. albicans*, *C. tropicalis* and *C. parapsilosis*, respectively).

While the number of antisense lncRNAs varied significantly across the three species, intergenic lncRNAs demonstrated somewhat similar distribution among the three CTG clade species (1457-1581), with *C. glabrata* having a lower number (444). Of note, when analyzing only the S dataset we also observed that *C. glabrata* had overall fewer intergenic lncRNAs (267) compared to the other species (1223, 943 and 794 for *C.*

albicans, *C. tropicalis* and *C. parapsilosis*, respectively). This observed uniform distribution of intergenic transcripts in CTG clade species and the larger difference with *C. glabrata* resembles the phylogenetic relationships of the studied species (Gabaldón, Naranjo-Ortíz and Marcet-Houben, 2016). Saturation plots for intergenic lncRNAs showed the numbers of lncRNAs obtained by subsets of 30-40 samples are similar to the ones obtained by the total datasets, suggesting that the obtained catalogues for intergenic lncRNA are very comprehensive. Of note, an effect of reduction of the total number of intergenic lncRNAs is observed in *C. albicans* as more dataset are analyzed, likely resulting from the fusion of previously fragmented predictions.

We then assessed several major features of lncRNAs and compared them to those of protein coding genes. Consistent with lncRNAs studies in other organisms (Lopez-Ezquerra, Harrison and Bornberg-Bauer, 2017; Pegueroles, Iraola-Guzmán, *et al.*, 2019), we found that both types of lncRNAs tend to be shorter and have lower GC content and levels of expression than protein coding genes (Fig. 5.2B, C, and D). We also observed that intergenic lncRNAs are longer than antisense transcripts (Wilcoxon rank sum test p-value < 0.05).

Identification of related lncRNAs across fungal species

We next explored the evolutionary relationships of lncRNAs across the studied species. Considering that lncRNAs generally show low levels of sequence conservation and may adopt conserved secondary structures, we used several independent methods to establish their relatedness. This is specifically relevant in the context of the current study, as the considered *Candida* species are photogenically very diverse.

Considering that parts of the antisense lncRNAs that overlap with known protein coding genes can potentially bias evolutionary inference, we performed these analyses only on the intergenic lncRNAs. We first identified one-to-one reciprocal BLAST best hits between each pair of species (see Materials and Methods for details), and used a clustering methodology developed in Pegueroles *et al.* (2019) to unify all pairwise species comparisons and define lncRNA families across species. As expected, this analysis identified a very small number of conserved lncRNAs families (Fig. 5.3). We observed only two families shared between *C. parapsilosis* and *C. tropicalis*, which in total contain four lncRNA genes. These results highlight the overall low sequence conservation of lncRNAs, and the high levels of species divergence.

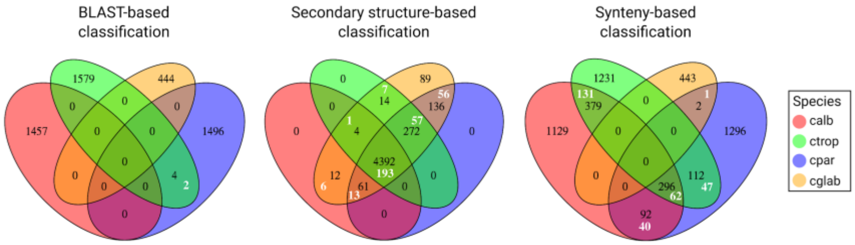


Fig. 5.3. Assessment of evolutionary relationships of intergenic lncRNAs across *Candida* species. In Venn diagrams, black numbers denote the number of lncRNAs shared across species and white numbers denote the number of clustering families.

We then analyzed the relatedness of lncRNAs based on their secondary structures (see Materials and Methods for details). In stark contrast to BLAST results, we observed that the majority of lncRNAs can be classified into structural families (~98%, Fig. 5.3). Moreover, most of the families are formed between all four species (88%), with only *C. glabrata* having species-specific lncRNAs structures (n=89). These results indicate that lncRNAs of the studied species share many secondary structure folds, despite the low level of sequence divergence, although it is unclear whether this similarity is the result of higher evolutionary constraints in the structure (Johnsson *et al.*, 2014; Jones and Sattler, 2019), or whether it can be attributed to a low specificity of the structural comparison approach.

We also searched for syntenic lncRNAs using a methodology developed and validated in Pegueroles *et al.* (2019) (see Materials and Methods for details). This analysis revealed a significantly higher number of evolutionary related lncRNA (n=881, 17.7%) than BLAST-based approach, but considerably lower than in case of secondary structures (Fig. 5.3). As expected, the majority of the syntenic relationships were observed between species of the CTG clade. In particular, the largest number of syntenic genes was found between a pair of *C. albicans* and *C. tropicalis* (379 lncRNAs groups in 131 syntenic families), followed by the triplet of the species (296 genes in 62 families) and the pair of *C. parapsilosis* and *C. tropicalis* (112 lncRNAs in 47 families). For the distantly related *C. glabrata* only 1 pair of syntenic lncRNA genes with *C. parapsilosis* was found.

Co-expression analysis

To further explore the molecular properties of lncRNAs in *Candida* pathogens and obtain functional insights, we carried out a gene co-expression analysis using the WGCNA approach (see Materials and

Methods). As in the case of evolutionary inference described above, we restricted the co-expression analysis only to intergenic lncRNAs.

After inspecting principal component analysis plots (suppl. Fig. S5.2) to remove outliers, and filtering out lowly expressed genes (TPM<0.1 in more than 80% of samples), we obtained sufficient power values (β =12-16) to generate scale-free co-expression network topologies for each of the species (suppl. Fig. S5.4).

For all species, a co-expression network analysis identified multiple highly interconnected modules (n=9-19, Fig. 5.4, suppl. Fig. S5.4). Interestingly, lncRNAs were present in the majority of modules (Fig. 5.4), with the only exception of *C. albicans*. This result highlights that lncRNAs are ubiquitous members of co-expressed gene clusters in all studied fungal pathogens. We also identified that two lncRNAs (1 for *C. tropicalis* and 1 for *C. glabrata*) represent the most highly connected nodes (hubs) in two modules (Fig. 5.4).

We further assessed the connectivities of lncRNAs in networks and compared them with those of protein coding genes. This analysis revealed that, despite being widely distributed across networks, lncRNAs have significantly lower total modular connectivities (Wilcoxon rank sum test p-value < 0.05) compared with coding genes.

To gain functional insights of lncRNAs involved in modules, we performed GO term enrichment analysis of protein coding genes in all modules (see suppl. files 1-4 for all GO terms of each module). This analysis showed a wide variety of GO term enrichment categories across all the modules, including GO terms related to fungal pathogenicity, such as “adhesion to symbiont” for *C. albicans*, and “filamentous growth” for C. *tropicalis.*

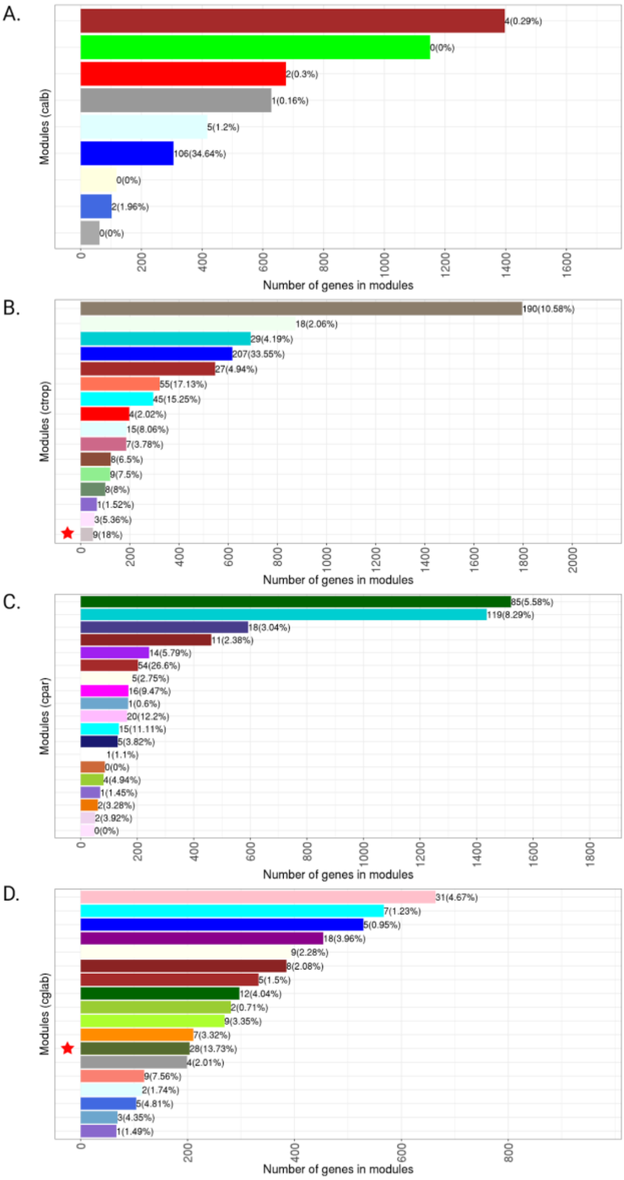


Fig. 5.4. Co-expressed modules and distribution of lncRNAs in modules for studied *Candida* species. Modules for **(A)** *C. albicans*; **(B)** *C. tropicalis*; **(C)** *C. parapsilosis* and **(D)** *C. glabrata*. Each barplot represents a module, which height corresponds to the number of genes involved in the module. Numbers at the right of bars represent the number of lncRNAs of the module and proportion (in parentheses) of lncRNA over the total number of genes in the module. Modules where a lncRNA is a hub are highlighted by red stars.

Expression dynamics of lncRNAs during course of epithelial infection

To investigate in more detail the possible implication of lncRNAs in *Candida* virulence, we resorted to the S dataset, which comprises expression data through the time-course of epithelial infection of the studied pathogens. To this end we performed a differential expression analysis to identify lncRNAs (both intergenic and antisense) which are deregulated throughout the infection. As shown in Fig. 5.5A, there are multiple lncRNAs (80-450) that significantly change their expression upon interaction with epithelial cells. This is observed from the initial time points after infection, with the number of differentially expressed lncRNAs increasing as infection progresses. Of note, the overall pattern of lncRNA deregulation in *C. glabrata* is somewhat lower compared with other species.

The 24h control samples present in the S dataset allowed us to identify lncRNAs which are differentially expressed in the absence of human epithelial cells, accounting for the effect of time and changes in growth medium. We observed that the process of infection and normal growth in culture medium deregulate largely overlapping sets of lncRNAs (Fig. 5.5B). This phenomenon was also observed for protein coding genes of these species (Liu et al., 2015, and Chapter 4), indicating that most non-coding and coding genes related to pathogenesis of these species are also related to standard growth metabolism. We nevertheless identified a substantial number of “infection-specific” lncRNAs, which are differentially expressed exclusively due to the infection process (n=189-273, depending on the *Candida* species).

Further analysis of the “infection-specific” lncRNAs showed no significant differences between their network connectivity compared with other lncRNAs. Additionally, we observed that a large portion of “infection-specific” intergenic lncRNAs are involved in modules (~19-49%, or ~0.1-6.3% from the total number of intergenic lncRNAs depending on species).

We then assessed syntenic relationships of “infection-specific” lncRNAs across species (suppl. Fig. S5.5), which showed that there are only two shared families between *C. albicans* and *C. tropicalis* (namely family147 and family51) containing syntenic “infection-specific” lncRNAs. Interestingly, two *C. tropicalis* lncRNAs, namely MSTRG.74.1 (in family147) and MSTRG.8910.1 (in family51) are involved in modules with GO terms “interaction with host” and “filamentous growth”, respectively. Overall, this analysis suggests that infection-specific genes tend to be specific for each *Candida* pathogen, in agreement with our previous results from protein-coding genes (see Chapter 4).

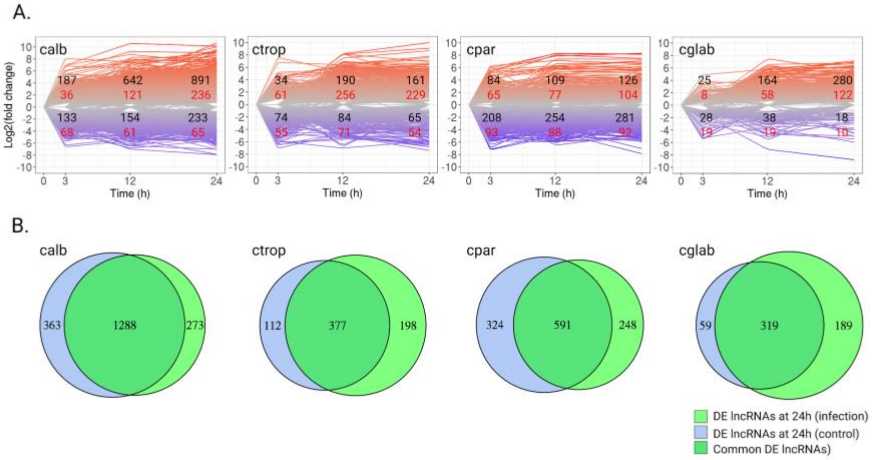


Fig. 5.5. lncRNAs during fungal infection of epithelial cells. **(A)** lncRNAs expression dynamics plots based on log2 fold changes compared to time point 0. Each line (up-regulated - red, down-regulated - blue) corresponds to expression levels of a lncRNA. Numbers on the plots indicate the number of differentially expressed lncRNAs (in red - intergenic lncRNAs, in black - antisense lncRNAs); **(B)** Venn diagrams of differentially expressed (padj<0.01) lncRNAs during infection (in green) and in control samples (in blue). The numbers of “infection-specific” lncRNAs are indicated in green-only portions of Venn diagrams.

5.5 Concluding remarks

Despite intensive research on lncRNAs during the recent decade (Jarroux, Morillon and Pinskaya, 2017), these enigmatic transcripts have never been systematically investigated in human fungal pathogens from the *Candida* clade. Here, we analyzed all RNA-Seq samples available for the studied yeasts, and reported the first comprehensive catalogs of the four major *Candida* pathogens. As in other species across a wide range of taxa, lncRNAs are abundant in these *Candida* yeasts. We classified the identified lncRNAs into intergenic and antisense transcripts, and show that antisense transcription of lncRNAs is widespread in *Candida*.

We identified that the major properties of *Candida* lncRNAs follow the same trends as lncRNAs in other species, i.e. they are shorter, have lower GC content and lower expression levels when compared to protein-coding genes. From the evolutionary standpoint, lncRNAs in *Candida* show poor primary sequence conservation, detectable level of synteny between CTG clade members and high secondary structure conservation. In fact, these properties of lncRNAs seem to be universal across plants, animals and fungi

(Lopez-Ezquerra, Harrison and Bornberg-Bauer, 2017; Zhao *et al.*, 2018; Borgognone *et al.*, 2019; Pegueroles, Iraola-Guzmán, *et al.*, 2019), hinting to their common evolutionary constraints. Moreover, co-expression network analysis revealed that lncRNAs of *Candida* are ubiquitously co-expressed with protein coding genes. Considering that highly co-expressed features are likely to be functionally related (Serin *et al.*, 2016), we show that lncRNAs can have numerous functional implications in *Candida*, with some of them related to fungal virulence. We then directly investigated the roles of the predicted lncRNAs in fungal virulence by assessing their expression dynamics during the course of epithelial infection of each *Candida* species. For each fungus, we identified a large number of infection-specific lncRNAs which are differentially expressed exclusively due to interaction with the human host. In fact, these transcripts can be considered as direct targets for further experimental analysis. Altogether, the lncRNAs catalogues inferred here serve as a valuable resource opening novel avenues for further in depth lncRNA research in human yeast pathogens.

5.6 Supplementary figures

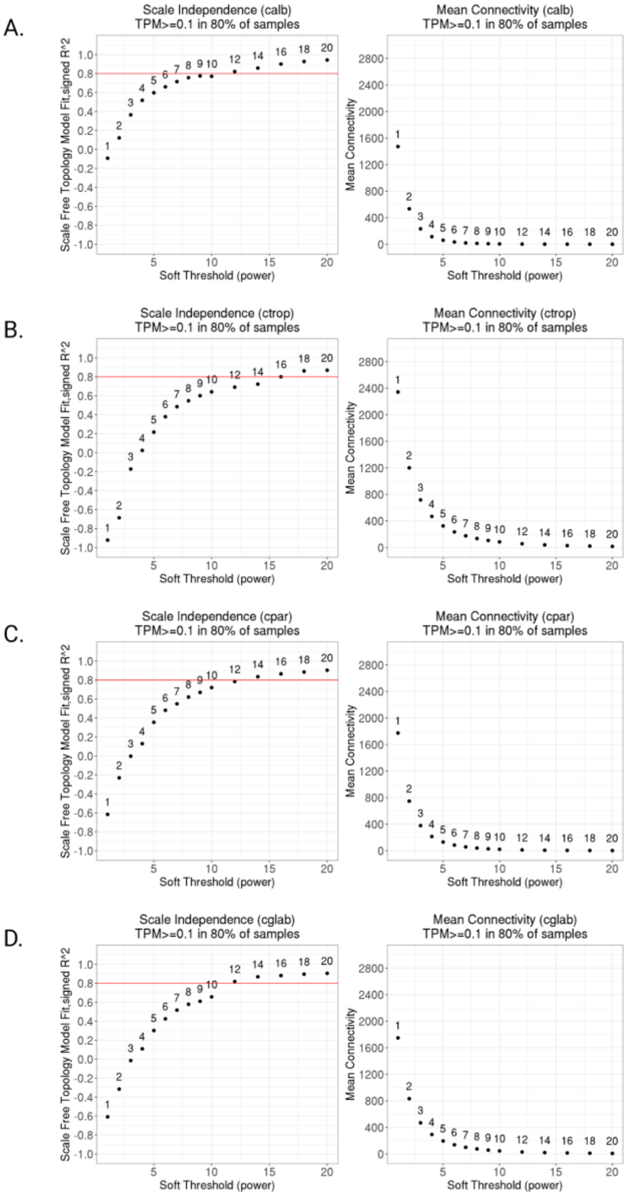


Fig. S5.1. Distribution of β power values for scale-free network topology (on the left) and mean network connectivity (on the right) when using TPM filtering and discarding outlier samples (see Materials and Methods for details). Plots for **(A)** *C. albicans*; **(B)** *C. tropicalis*; **(C)** *C. parapsilosis* and **(D)** *C. glabrata*.

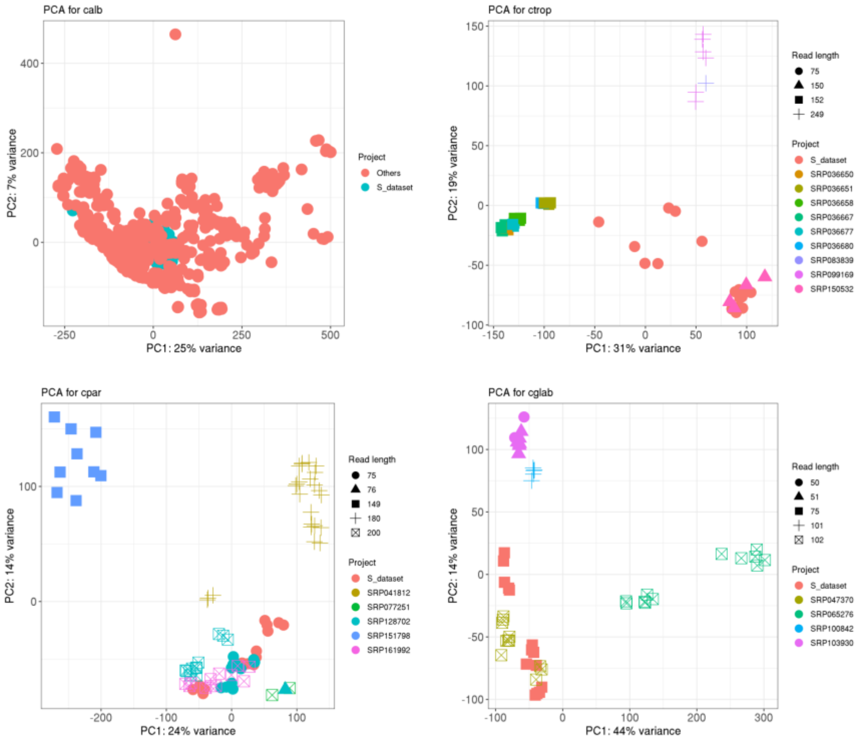


Fig. S5.2. PCA plots of all analysed samples across species (top-left corner of each plot). “calb” denotes *C. albicans*, “cglab” - *C. glabrata*, “cpar” - *C. parapsilosis* and “ctrop” - *C. tropicalis*. The plots are generated using log-transformed TPM expression values. Color codes correspond to SRA project accession numbers and “S” dataset used in this study.

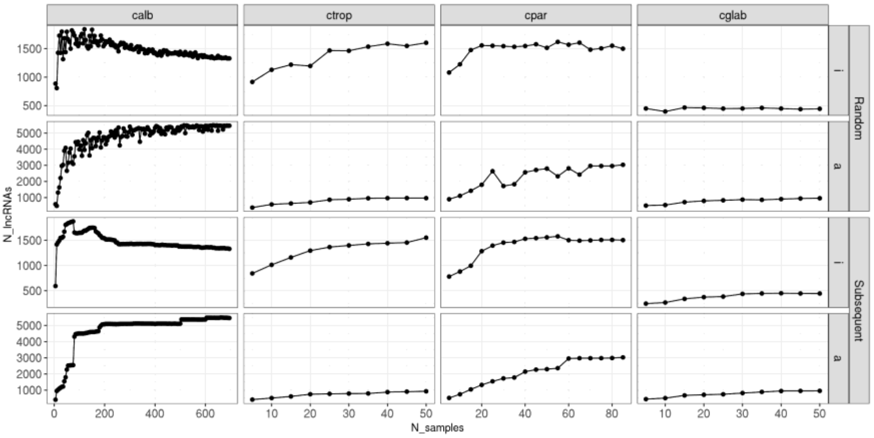


Fig. S5.3. Saturation plots showing the number of identified lncRNAs depending

on the number of analyzed samples. Per each species, samples were chosen either randomly or subsequently, and the number of analyzed samples was incremented by 5 at each step.

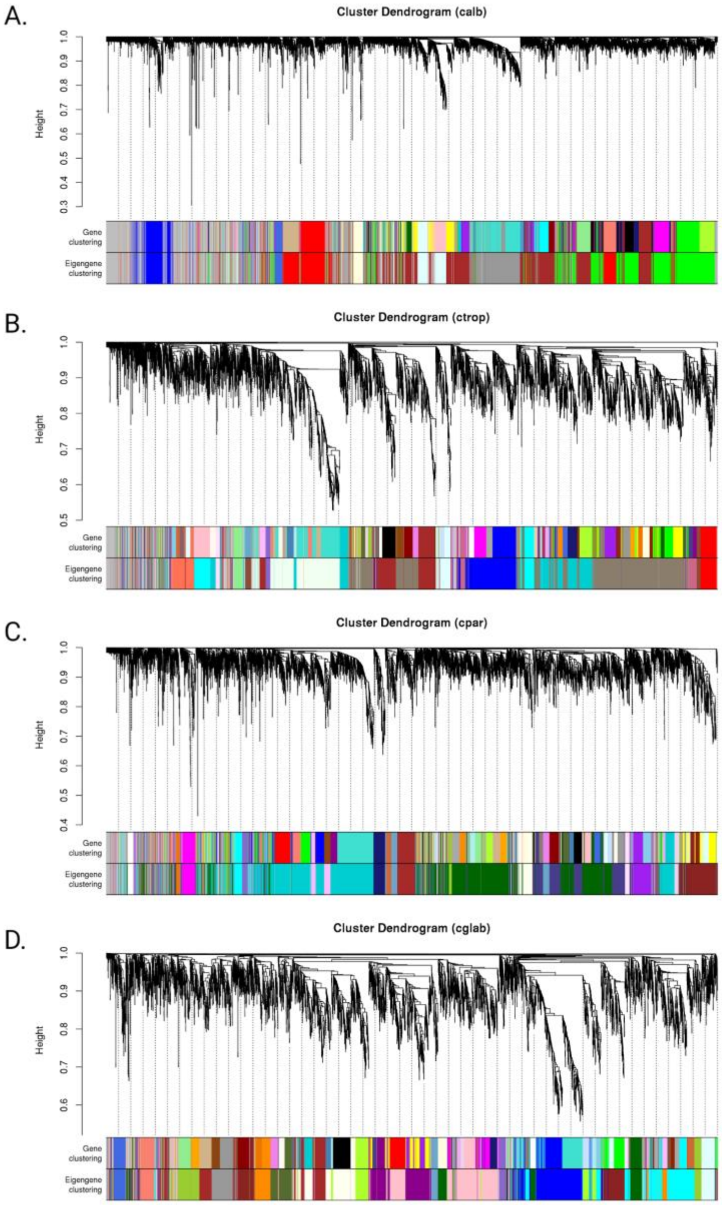


Fig. S5.4. Gene co-expression networks represented as dendrograms produced with WGCNA and *hclust* function based on 1-Topology Overlap Matrix. Networks for **(A)** *C. albicans*; **(B)** *C. tropicalis*; **(C)** *C. parapsilosis* and **(D)** *C. glabrata*. Each

dendrogram represents co-expressed modules (on the bottom) obtained by gene clustering (top row) and eigengene clustering (bottom row).

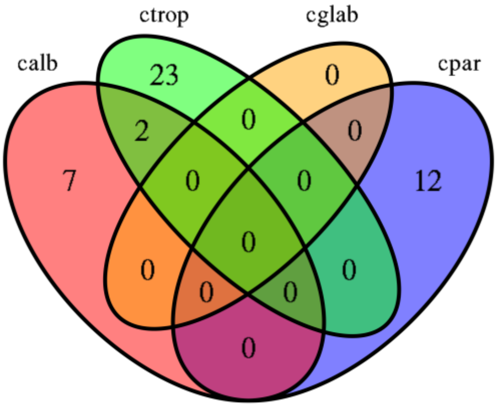


Fig. S5.5: Syntenic families where "infection-specific" lncRNAs are involved in. "calb" denotes *C. albicans*, "cglab" - *C. glabrata*, "cpar" - *C. parapsilosis* and "ctrop" - *C. tropicalis.*

6 Probe-based enrichment of fungal transcriptomes from human-derived samples enables *in vivo* study of *Candida* spp. infections

6.1 Abstract

The study of transcriptomic interactions between host and pathogens in *in vivo* conditions are frequently hampered by the low relative amount of RNA from the pathogen. Yeast opportunistic pathogens of the genus *Candida* pose an important medical problem, and cause superficial mucosal infections as well as life-threatening systemic infections in susceptible patients. Four major phylogenetically diverse species account for over 90% of *Candida* infections, and their specific interactions with various human tissues are still poorly understood. Here, we designed and validated pan-*Candida* target capture probes to enrich protein-coding transcripts and long non-coding RNAs of these four major *Candida* pathogens. Our approach outperformed enrichment based on differential lysis of host cells, and showed similar enrichment performance but better fidelity of expression levels than previous target capture designs, and moreover enabled species multiplexing and lncRNA targeting. In addition, we show that our probe-based enrichment strategy allows genotyping of the infecting strain and qualitative assessment of the microbiota present in the sample.

Key words: *Candida*, host-pathogen interactions *in vivo*, RNA-Seq, probe-based enrichment

6.2 Introduction

Human fungal pathogens pose a serious global healthcare problem. The incidence of fungal infections, in their various forms, has increased over the last decade (Oren and Paul, 2014), currently affecting 25% of the global population and causing 1.5 million deaths every year (Havlickova, Czaika and Friedrich, 2008; Bongomin *et al.*, 2017). *Candida* yeasts are the most common cause of invasive fungal infections (Neofytos *et al.*, 2013; Guinea, 2014). Current challenges to overcome *Candida* infections include the difficulty of accurate diagnosis, due to the high genetic diversity of this group, and the emergence of novel pathogenic species (Satoh *et al.*, 2009; Papon *et al.*, 2013; Gabaldón, Naranjo-Ortíz and Marcet-Houben, 2016; Consortium OPATHY and Gabaldón, 2019). Additionally, therapeutic options are limited and are further losing efficiency due to the high capacity of adaptation to antifungal drugs shown by some *Candida* species (Cortegiani *et al.*, 2018; Ksiezopolska and Gabaldón, 2018).

There are over 30 different *Candida* species that can infect humans (Papon *et al.*, 2013). However, more than 90% of the infections are caused by the four species with the largest global incidence: *C. albicans*, *C. glabrata*, *C. parapsilosis* and *C. tropicalis* (Guinea, 2014). Despite their common genus name, these yeasts are phylogenetically diverse and have close non-pathogenic relatives, which suggests that their pathogenicity towards humans has emerged independently (Gabaldón, Naranjo-Ortíz and Marcet-Houben, 2016). Given the likely differential nature of their infections, there is a need to better understand virulence mechanisms and host-pathogen interactions specifically for each of these diverse species. However, the majority of studies so far have focused on *C. albicans* (Mayer, Wilson and Hube, 2013), with research on non-albicans species significantly lagging behind. As a result, we still have a very poor understanding of how virulence and host-pathogen interplay vary across *Candida* pathogens.

Considering that host-fungus interactions comprise mutual adaptive processes involving dynamic changes in molecular pathways of both organisms, studies addressing gene regulation and expression during infection can be very informative. Recent advances in Next Generation Sequencing (NGS) allow studying molecular interactions by assessing transcriptome dynamics of host and pathogen simultaneously (Amorim-Vaz and Sanglard, 2015; Enguita *et al.*, 2016; Westermann and Vogel, 2018; Wolf *et al.*, 2018; Hovhannisyan and Gabaldón, 2019). In this regard, high throughput transcriptome sequencing (RNA-Seq) has proven to be a powerful method for disentangling molecular interactions between human and *Candida* pathogens (Amorim-Vaz *et al.*, 2015; Bruno *et al.*, 2015; Liu *et al.*, 2015; Rasheed, Battu and Kaur, 2018). However, a major limitation of sequencing-based transcriptomic approaches, particularly when used *invivo*, is the very low ratio of fungal/host RNA, which severely limits downstream analyses. For instance, in a recent RNA-Seq study using murine model of vaginal candidiasis, the fungal reads represented less than 0.1% of the total reads (Bruno *et al.*, 2015).

There are two main approaches for overcoming this obstacle. The first, indirect way, is to study host pathogen interactions *in vitro* or by using different animal models, whereby fungal loads can be controlled (MacCallum, 2012; Segal and Frenkel, 2018). In fact, the majority of studies investigating pathogenicity mechanisms of *Candida* pathogens have been performed *in vitro* (Enjalbert, Nantel and Whiteway, 2003; García-Sánchez *et al.*, 2004). However, it has been shown that, for example, *C. albicans* displays different transcriptional dynamics *in vivo* and *in vitro*, indicating that *in vitro* conditions only partially reflect real infections (Xu *et al.*, 2015). Furthermore, while animal models (especially mammals) better resemble the human host, there are economic and ethical issues precluding their extensive usage. A second approach consists of enriching the fungal component from the mixed host-fungus sample. This enrichment can be performed by physical methods or at the molecular level. In the first case, fungal cells in the sample can be selected, for example, by selectively lysing the host cells (Rodríguez *et al.*, 2019). In the second case, enrichment is achieved by selecting fungal RNA/DNA molecules by, for example, using oligonucleotide probes specifically hybridizing to fungal genetic material. As such, SureSelect probe-based enrichment technology has been applied to enrich *C. albicans* transcripts from murine and *Galleria mellonella* infection models to subsequently perform RNA-Seq (Amorim-Vaz *et al.*, 2015). That study showed that probe-based enrichment increased fungal RNA relative abundance up to 1600-fold, while only significantly altering the expression levels of ~3% of the genes. This technology was also successfully used on a mouse model of aspergillosis (Chung *et al.*, 2018). Host cell lysis approach has also shown promising results. A recent study (Rodríguez *et al.*, 2019) evaluated several ways of selectively lysing human cells, and showed that a specific lytic treatment with Buffer LRT followed by centrifugation effectively retained *C. albicans* cells from a mixture with human cells, while preserving fungal genetic material with high quality. However, that study did not evaluate the applicability of this strategy with RNA-Seq.

Here, we designed a novel target-capture enrichment approach based on SeqCap (Roche) technology that improves over the previously existing enrichment designs in two main ways. First, in addition to protein-coding genes, our probe set also targets long-non coding RNAs (not the ones reported in Chapter 5), enabling the analysis of these poorly studied molecules. Secondly, it combines targets for the full transcriptomes (coding and non-coding) of the four main *Candida* pathogens, expanding its use for comparative transcriptomic analyses of different species, the study of co-infections, and the direct study of clinical specimens from the majority of cases of candidiasis. We tested the efficacy of our enrichment approach and the accuracy of downstream RNA-Seq analyses by using human vaginal swab samples spiked with defined numbers of *C. albicans* cells. Additionally, we compared the results of targeted enrichment with the aforementioned enrichment method based on selective lysis of human cells (Rodríguez *et al.*, 2019). Our results indicate that our approach efficiently enriched fungal RNAs and did so to a significantly higher level as compared with the differential lysis approach. Most importantly, the probe-based enrichment did not significantly alter expression levels, with only ~0.3-1.5% of the fungal genes being affected. Moreover, we show that besides standard RNA-Seq analyses such as transcriptome profiling and differential gene expression, the targeted enrichment results can serve to perform additional analyses such as variant calling and meta-transcriptome profiling.

6.3 Materials and Methods

Preparation of spiked-in vaginal samples

A total of 48 *Candida*-negative vaginal swabs (E-Swabs, COPAN Diagnostics, CA, USA), from 48 pre-menopausal, non-pregnant and healthy women of at least 18 years of age, were collected at the University Hospital of Ghent (Ghent, Belgium, informed consent was obtained from all participants. Approval ethical committee EC/2016/0192). Upon collection, the samples were immediately immersed in 1.2 ml RNAlater (ThermoFisher Scientific, Waltham, MA) and stored first at 4 ºC overnight and then at -80 ºC. Absence of *Candida* was tested by a) culture in Sabouraud glucose agar with chloramphenicol (Merck KGaA) and subsequent species screening with MALDI-TOF analysis, b) microscopic visualization through wet mount in combination with phase contrast microscopy and c) Gram-staining in combination with light microscopy.

For subsequent analysis, stored vaginal samples were thawed at 37 ºC and briefly vortexed. Swabs were then discarded and the RNAlater solutions containing the cells were all pooled together in a 100 ml falcon tube and mixed by pipetting. Then, two aliquots of 12.5 ml of the vaginal samples pool were spiked with *Candida albicans* SC5314 cells to reach a final concentration of 10_5 cells/ml, and 10_3 cells/ml respectively, and then split in 1-ml aliquots and stored at -80 ºC until RNA extraction. Prior to preparation of spike-in samples, *C. albicans* had been incubated at 32 ºC in Sabouraud plates overnight, whereafter a colony was incubated in YPD broth at 32 ºC, with shaking, overnight. Fungal cells were counted with a microscope by using a hemocytometer (Neubauer chamber).

Enrichment with buffer RLT + β-mercaptoethanol of vaginal samples

Because the use of buffer RLT + β-mercaptoethanol enrichment (treatment "B") might change gene expression, which would no longer match the transcriptional profile encountered during the infection, we did not only test the efficiency of the enrichment in vaginal samples but also checked whether gene expression is affected during the enrichment process. To this purpose, we included a control in which RNA extraction was performed directly from the vaginal samples without a previous lytic enrichment. In addition, we also included an enrichment with pre-treatment with thiolutin (treatment "BT"), a known inhibitor of transcription (Jimenez, Tipper and Davies, 1973; Tipper, 1973; Kebaara *et al.*, 2006; Pelechano and Pérez-Ortín, 2008) to prevent gene expression changes. Finally, we also used thiolutin without enrichment to test whether the thiolutin modified gene expression (treatment “T”). In summary, four different treatments were tested: 1) No thiolutin and no lytic enrichment (Treatment “N”); 2) Thiolutin and no lytic enrichment (Treatment “T”); 3) No thiolutin and buffer RLT + β-mercaptoethanol lytic enrichment (Treatment “B”); 4) Thiolutin and Buffer RLT + β-mercaptoethanol lytic enrichment (Treatment “BT”).

All treatments were performed in triplicate, starting from 1- ml aliquots, which had been stored at -80 ºC, of a pool of vaginal samples spiked with either 10_5 or 10_3 *Candida* cells. One-ml aliquots were thawed and centrifuged at maximum speed (> 20 000 *g*) for 5 min in a benchtop centrifuge to collect human and fungal cells and to discard RNAlater. Pellets of cells were then used for treatments N, T, B or BT. A total of six vaginal samples per run, three containing 10_5 *Candida* cells and three with 10_3 *Candida* cells, were used for each different treatment. Each run consisted of the particular treatment N, T, B or BT followed by RNA extraction with the RiboPure Yeast Kit (ThermoFisher Scientific), performed following the manufacturer’s instructions. We carried out only six extractions per run to minimize handling time in order to minimize post-sampling changes in expression and RNA degradation.

For treatment N, pellets containing human and fungal cells were directly used for RNA extraction with the RiboPure Yeast Kit.

For treatment T, pellets of human and fungal cells were resuspended in ice-cold PBS with thiolutin and incubated on ice for 15 min as shown in treatment BT. Pellets of cells after centrifugation at 20 000 *g* for 8 min, at 4 ºC, were directly used for RNA extraction.

For treatment B, pellets of human and fungal cells were resuspended in 600 µl of RLT buffer containing 1% β-mercaptoethanol (*i.e.*, 143 mM) and pipetting up and down to lyse human cells. Samples were centrifuge at max speed (> 20 000 *g*) for 8 min, at 4 ºC, to collect intact yeast cells. Supernatants containing cell debris and nucleic acids from human cells were carefully discarded without disturbing the fungal cell pellet.

For treatment BT, pellets of human and fungal cells were resuspended in 200 µl ice-cold PBS containing thiolutin at a final concentration of 20 µg/ml. Samples were incubated for 15 min on ice to stop transcription events. Then, samples were centrifuged at 20 000 *g* for 8 min, at 4 ºC, to collect the cells. Pellets followed treatment B prior to RNA extraction, *i.e.*, RLT buffer with β-mercaptoethanol was used to lyse human cells followed by centrifugation to pellet fungal yeast cells.

Custom design of oligonucleotide probe-based targeted enrichment

We designed a custom pan-*Candida* enrichment kit for SeqCap technology (Roche, Basel, Switzerland). Our design included probes targeting whole transcriptomes of the four most widespread *Candida* pathogens - *C. albicans*, *C. glabrata*, *C. parapsilosis* and *C. tropicalis*, including annotated features, such as protein coding genes and non-coding RNAs, and newly annotated long non-coding RNAs (lncRNAs), which were predicted in our study (see below). For previously annotated features, we first obtained the corresponding reference genomes and genome annotations for each species: *C. albicans* SC5314 strain (assembly 22), *C. glabrata* CBS138, *C. parapsilosis* CDC317 and *C. tropicalis* MYA-3404 from the *Candida* Genome Database (last accessed in July 2017) (Skrzypek *et al.*, 2017). Considering that the genome of *C. albicans* is phased, we used only haplotype A for all downstream analyses. For each species, we extracted transcriptomes from genomes and genome annotations using the *getfasta* function of bedtools v. 2.26.0 (Quinlan, 2014).

We retrieved a predicted set of *C. parapsilosis* lncRNAs (n=618) from (Thuer, 2017). For the remaining species, we performed *de novo* prediction of lncRNAs. For this, we used all publicly available RNA-Seq datasets for these species available at the Sequence Read Archive (SRA) database as of June 2017 (Leinonen *et al.*, 2011). This comprised a total of 69 samples for *C. albicans*, 39 samples for *C. glabrata* and 36 samples for *C. tropicalis* (list of used sample accession numbers is available in suppl. table S1). Reads were mapped against the corresponding reference genome using TopHat2 v. 2.1.1 (Kim *et al.*, 2013). We then performed reference guided transcriptome assembly using Cufflinks v. 2.2.1 (Trapnell *et al.*, 2012), and, for each species, merged individual assemblies into a unified assembly and compared it with reference annotations. Subsequently, we selected novel intergenic transcripts longer than 200 bps and assessed their coding potential using CPC v. 0.9 software (Kong *et al.*, 2007). Transcripts with no coding potential were considered as lncRNAs. These analyses resulted in 187, 93 and 485 putative lncRNAs for *C. albicans*, *C. glabrata* and *C. tropicalis*, respectively. Additionally, we included the sequences of the External RNA Controls Consortium (i.e. ERCC) spike-in RNA control molecules to our probe design. Once all necessary sequences were obtained, we used a custom python script to design the probes targeting these sequences. This script designs probes with variable length (ranges between 55-70 bps), optimizing GC content, transcript coverage and number of probes per transcript. The custom probes were subsequently ordered from Roche and received as a SeqCap RNA Developer Enrichment Kit.

Library preparation

Sequencing libraries were prepared using the TruSeq Stranded mRNA Sample Prep Kit v2 (ref. RS-122-2101/2, Illumina) according to the manufacturer's protocol for all samples. All reagents subsequently mentioned are from the TruSeq Stranded mRNA Sample Prep Kit v2 if not specified otherwise. 500 ng of total RNA were used for poly(A)-mRNA selection using streptavidin-coated magnetic beads. Briefly, all samples were subsequently fragmented to approximately 300bp. cDNA was synthesized using reverse transcriptase (SuperScript II, Invitrogen) and random primers. The second strand of the cDNA incorporated dUTP in place of dTTP. Double-stranded DNA was further used for library preparation. dsDNA was subjected to A-tailing and ligation of the barcoded Truseq adapters. All purification steps were performed using AMPure XP Beads (Agencourt). Library amplification was performed by PCR on the size selected fragments using the primer cocktail supplied in the kit. Final libraries were analyzed using Agilent DNA 1000 chip (Agilent) to estimate the quantity and check size distribution, and were then quantified by qPCR using the KAPA Library Quantification Kit (KapaBiosystems) prior to amplification with Illumina’s cBot.

Before sequencing, 15 µl of each library was used to perform fungal RNA enrichment using a our custom SeqCap RNA Developer Enrichment Kit (see above for details on the kit design), following manufacturer’s instructions. Briefly, we first prepared the multiplex cDNA sample library pool (a mixture of all libraries), which was mixed together with 5 µg of COT Human DNA and 2,000 pmol of the corresponding multiplex hybridization enhancing oligo pool (to prevent hybridization between adapter sequences). After drying this mixture in a DNA vacuum concentrator at 60°C, the following reagents were added: 7.5 µl of 2X Hybridization Buffer and 3 µl of Hybridization Component A. Samples were vortexed for 10 seconds, centrifuged at maximum speed for 10 seconds, and then left at 95ºC for 10 min to denature the cDNA. After a short centrifugation at maximum speed for 10 seconds, the mixture was transferred to a 4.5 µl aliquot of SeqCap RNA probe pool previously prepared in a 0.2 ml PCR tube, vortexed for 3 seconds and centrifuged at maximum speed for 10 seconds more. Finally, the mixture was incubated in a thermocycler at 47°C for 20 hours (with the thermocycler lid set at 57°C). After the hybridization step, the sample was washed and the captured multiplex cDNA sample was recovered from the mixture with SeqCap streptavidin Beads, and amplified following the manufacturer's instructions. PCR products were purified and ready to use with AMPure XP Beads (Beckman Coulter).

The quality of the enriched pool was assessed with a Bioanalyzer DNA 1000 chip (Agilent). Non-enriched libraries and the enriched pool of libraries were loaded and sequenced using 2 x 125 red length on Illumina's HiSeq 2500.

DNA extraction, library preparation and sequencing of 16S rRNA

DNA was extracted from a pool of vaginal samples using the DNeasy PowerLyzer PowerSoil Kit (Qiagen, ref. QIA12855) following manufacturer's instructions. The extraction tube was agitated twice in a 96-well plate using Tissue lyser II (Qiagen) at 30 Hz/s for 5 min.

4 µl of the extracted DNA were used to amplify the V3–V4 regions of the bacterial 16S ribosomal RNA gene, using the following universal primers in a limited cycle PCR: V3-V4-Forward (5′-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG-3′) and V3-V4-Reverse (5′-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC-3′). To prevent unbalanced base composition in further MiSeq sequencing, we shifted sequencing phases by adding various number of bases (from 0 to 3) as spacers to both forward and reverse primers (we used a total of 4 forward and 4 reverse primers). The PCR was performed in 10 µl volume reactions with 0.2 µM primer concentration and using the Kapa HiFi HotStart Ready Mix (Roche, ref. KK2602). Cycling conditions were initial denaturation of 3 min at 95 °C followed by 20 cycles of 95 °C for 30 s, 55 °C for 30 s, and 72 °C for 30 s, ending with a final elongation step of 5 min at 72 °C.

After the first PCR step, water was added to a total volume of 50 µl and reactions were purified using AMPure XP beads (Beckman Coulter) with a 0.9X ratio according to manufacturer's instructions. PCR products were eluted from the magnetic beads with 32 µl of Buffer EB (Qiagen) and 30 µl of the eluate were transferred to a fresh 96-well plate. The primers used in the first PCR contain overhangs allowing the addition of full-length Nextera adapters with barcodes for multiplex sequencing in a second PCR step, resulting in sequencing ready libraries. To do so, 5 µl of the first amplification were used as template for the second PCR with Nextera XT v2 adaptor primers in a final volume of 50 µl using the same PCR mix and thermal profile as for the first PCR but only 8 cycles. After the second PCR, 25 µl of the final product was used for purification and normalization with SequalPrep normalization kit (Invitrogen), according to the manufacturer's protocol. Libraries were eluted in 20 µl and pooled for sequencing.

Final pools were quantified by qPCR using Kapa library quantification kit for Illumina Platforms (Kapa Biosystems) on an ABI 7900HT real-time cycler (Applied Biosystems). Sequencing was performed in Illumina MiSeq with 2 × 300 bp reads using v3 chemistry with a loading concentration of 18 pM. To increase the diversity of the sequences 10% of PhIX control libraries were spiked in.

Two bacterial mock communities were obtained from the BEI Resources of the Human Microbiome Project (HM-276D and HM-277D), each contained genomic DNA of ribosomal operons from 20 bacterial species (Willis *et al.*, 2018). Mock DNAs were amplified and sequenced in the same manner as vaginal sample. Negative controls of the DNA extraction and PCR amplification steps were also included in parallel, using the same conditions and reagents. These negative controls provided no visible band or quantifiable DNA amounts by Bioanalyzer, whereas our sample provided clearly visible bands after 20 cycles.

RNA-Seq analysis

We performed quality control of raw sequencing data using FastQC v. 0.11.6 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and Multiqc v. 1.0 (Ewels *et al.*, 2016). Samples having bases with low quality or traces of adapter sequences were filtered using Trimmomatic v. 0.36 with TruSeq3-PE-2 adapters with 2:30:10 options and minimum read length of 50 bps (Bolger, Lohse and Usadel, 2014). Filtered reads were mapped to the concatenated reference genomes of *C. albicans* SC5314 haplotype A (assembly 22) and primary human genome assembly GRCh38 obtained from Ensembl database (release 89, last accessed in June 2018) (Hunt *et al.*, 2018). Read mapping was performed using splice-junction aware aligner STAR v. 2.5.2b (Dobin *et al.*, 2013) using basic two-pass mode and default parameters. Read summarization was performed by STAR and Featurecounts v. 1.6.4 (Liao, Smyth and Shi, 2014). The counting of reads mapped to human and yeast genomes was performed with a custom python script read_count.py v. 1 available at https://github.com/Gabaldonlab. Reads that mapped to the two genomes equally well were discarded from the analysis.

Differential gene expression analysis was performed using the DESeq2 v. 1.26 Bioconductor package (Love, Huber and Anders, 2014). Genes with |log2 fold change| > 2 were considered as differentially expressed (DE) (Amorim-Vaz *et al.*, 2015).

To compare the fold enrichments resulting from our probe design and from Amorim-Vaz *et al.* (2015), we first modeled the dependency of fold enrichment from our initial fungal proportions on a log2 scale using the *lm* function of R v. 3.5.3. Subsequently, based on this model and using the *predict* function, we predicted the values of fold enrichment obtained with our probes in case these would have been applied to the initial fungal proportions observed in Amorim-Vaz *et al.* (2015).

Meta-transcriptomics analysis

Reads that were neither mapped to human nor to the yeast genome were subsequently used for meta-transcriptomic analysis using the Centrifuge v. 1.0.4 software (Kim *et al.*, 2016). The data were mapped to a database comprising genomes of all bacteria, archaea, fungi and man, available for *centrifuge-download* function as of August 2018. To compare the meta-transcriptomic results with actual metagenomic profiles of samples, we also performed a 16S rRNA-based metagenomic sequencing of one sample from the pool of negative vaginal specimens. The data were processed with DADA2 v. 1.14.1 Bioconductor package (Callahan *et al.*, no date).

Variant calling

We used the bam files obtained by STAR and performed variant calling using bcftools v. 1.6 with mpileup function, setting --max-depth 2000 option and default parameters. For downstream analysis we used variants with QUAL>20. To test whether these variants allow identification of the spiked *C. albicans* strain SC5314, we compared the variants called in our study with those of 58 strains representing the major clades of the *C. albicans* phylogeny (Ropars *et al.*, 2018). Variant calling of those samples was done as described previously (Mixão and Gabaldón, 2020). Considering that the data of previously published strains were generated on the basis of whole genome sequencing, we first performed a variant subsetting with vcftools v0.1.16 of all vcf files to analyze only variants within coding regions, and then retained the variants with QUAL>20. Further, each of our samples with N treatment (i.e. without lytic enrichment) was compared with 58 strains using the bcftools *isec* function.

All other data analyses and visualizations were performed in R v. 3.5.3 using various packages.

6.4 Results

Design of a pan-Candida enrichment kit targeting coding and non-coding transcriptomes

We set out to design and validate a probe-based enrichment approach suitable for the analysis of coding and non-coding transcriptomes of the four major *Candida* pathogens in samples with low content of fungal RNA, such as human-derived samples. For this, we designed probes targeting all annotated protein-coding genes in *C. albicans*, *C. glabrata*, *C. parapsilosis* and *C. tropicalis.* In addition, we predicted lncRNAs of these species and included probes for them in our design (see Materials and Methods). Our final dataset targets 24,282 previously annotated features, including protein coding genes and non-coding RNAs and 1,383 newly predicted lncRNAs (not the ones reported in Chapter 5) from the four mentioned species. We chose the SeqCap enrichment approach by Roche, although our design could easily be adapted to other technologies. In brief, SeqCap is a hybridization-based technology, which uses biotinylated oligonucleotide probes that specifically bind to NGS libraries of interest prior to sequencing. After target binding, probes hybridized with libraries are pulled by streptavidin-coated magnetic beads to obtain sequencing libraries, which are highly enriched with target regions.

Targeted probe-based capturing efficiently enriches fungal transcripts and outperforms differential lysis approaches

To test our targeted enrichment approach and compare it with the lytic enrichment, we used both approaches in parallel and in different combinations on a control sample consisting of a pool of vaginal swabs from healthy donors spiked with a known number of *C. albicans* cells (Fig. 6.1, see Materials and Methods). After lytic enrichment, all samples were subjected to RNA extraction and RNA-Seq library preparation. All libraries were split into two equal parts, one of which was further enriched with SeqCap oligonucleotide probes. Finally, all obtained samples were subjected to RNA-Seq in triplicates.

After RNA-Seq, we mapped the data to a concatenated *C. albicans* and human reference genome, and calculated proportions of reads mapped to either yeast or human (Fig. 6.2). Our results show that probe-based enrichment significantly increases the fraction of fungal reads as compared to non-enriched samples.

Fold enrichment was proportionally higher in samples having lower initial amounts of fungal RNA - from 70.8 to 99% (1.4 fold enrichment) for 10_5 fungal cells/ml of vaginal sample and from 4.8% to 85.4% (~17 fold enrichment) for 10_3 cells, on average across replicates . Second, we show that the differential lysis approach does not enrich fungal data significantly (with an exception of BT at low fungal load), when enrichment over the bulk transcriptome is assessed.

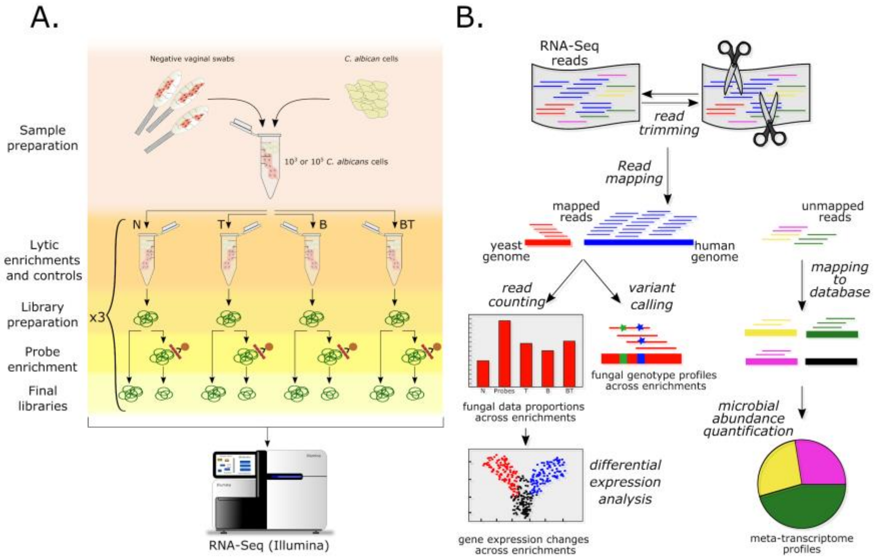


Fig. 6.1. (A) Schematic representation of the experimental setup. A pool of *Candida*-negative vaginal samples was spiked with different loads of *C. albicans* cells. Spiked samples further underwent different enrichments and sequencing (Materials and Methods for details of N, T, B and BT lytic enrichments); **(B)** Schematic RNA-Seq data analysis workflow employed in this study (see Materials and Methods for details).

Overall, our data suggests that targeted probe enrichment significantly outperforms human cell lysis method in enriching fungal RNA from clinical samples. We also assessed the efficiency of the enrichment strategies for the set of newly predicted lncRNAs (suppl. Fig. S6.1). As expected, we obtained proportionally similar enrichment as for the whole dataset (Fig. 6.2). Of note, with regards to lncRNAs, B method had similar enrichment efficiency as probe-based enrichment.

Further, we aimed to compare the efficacy of our targeted enrichment with the results of the study of Amorim-Vaz *et al.* (Amorim-Vaz et al. 2015). This study has reported significantly higher fold enrichment, ranging from 670 to 1670, than observed in our case (maximum of 43). However, that study was performed using mouse and *G. mellonella* animal models, whereby the initial fungal proportions of 0.03%-0.1% were significantly lower than in our study. In fact, the highest initial fungal proportion observed in that study (0.1%) is 15 fold lower than the lowest proportion in our experiment (~1.48%). To make our results comparable, we modeled the dependency of fold enrichment on the initial fungal proportion. Then we predicted the fold enrichment values which potentially could be obtained with our probes if using the same initial proportions of fungal RNA as in the study of Amorim-Vaz *et al.* (2015). The results of this modeling (suppl.

Fig. S6.2) show that the actual fold enrichment of our probes and of those used in the previous report are similar.

High fidelity of gene expression levels after probe-based enrichment

We further tested whether probe-based enrichment alters the expression levels of fungal genes, which could bias downstream analyses. First, we analyzed the mean normalized read count data of the same samples before and after probe-based enrichment (Fig. 6.3A).

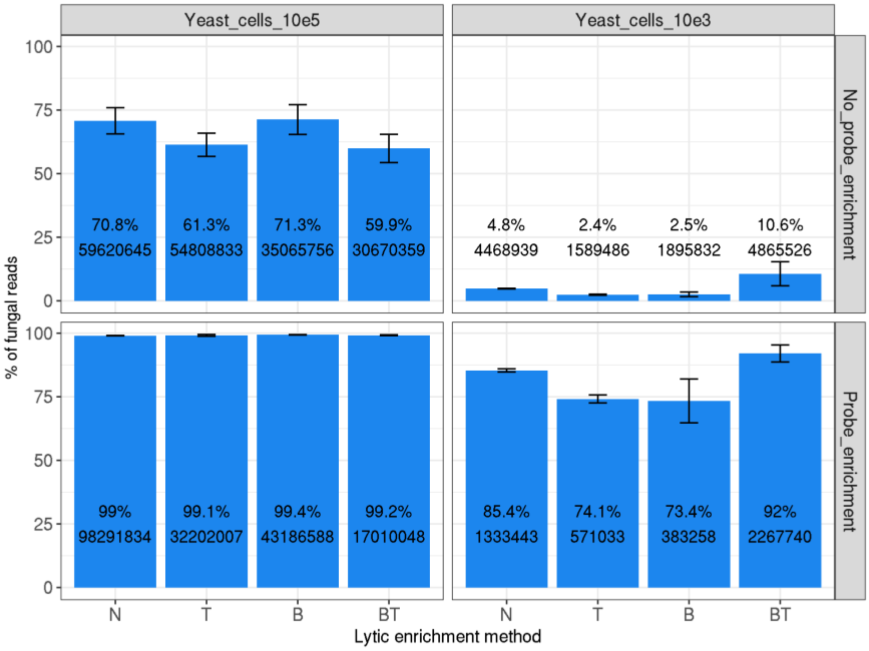


Fig. 6.2. Proportions of fungal RNA reads obtained with different enrichment approaches, as determined by RNA-Seq.

Differential gene expression analysis showed that targeted enrichment significantly alters the expression of a small number of genes (n=23-104), constituting ~0.3-1.5% of the *C. albicans* transcriptome. The proportion of genes with altered expression in our study is lower than that reported in the study of Amorim-Vaz *et al.* (2015), i.e. ~3%. Reproducing the results of that report and applying the same analysis approaches for the data of both studies indicates that our targeted enrichment approach preserves the true expression levels more efficiently (suppl. Fig. S6.3).

Notably, most of the DE genes in our study have higher expression in non-probe enriched samples and are highly expressed, particularly in the samples with higher fungal load. These two observations suggest that the probes targeting these genes reach saturation. We further assessed whether genes with altered expression were randomly distributed or common between analysed conditions.Taking advantage that our experimental design allowed performing differential expression analysis between all pairs of original and probe-enriched samples in all tested conditions, we were able to show that most of the genes with biased expression are common between conditions (Fig. 6.3B). This observation indicates that oligonucleotide probes systematically bias the expression levels of the same genes (see suppl. table S1), allowing us to confidently identify and discard these genes from further analyses.

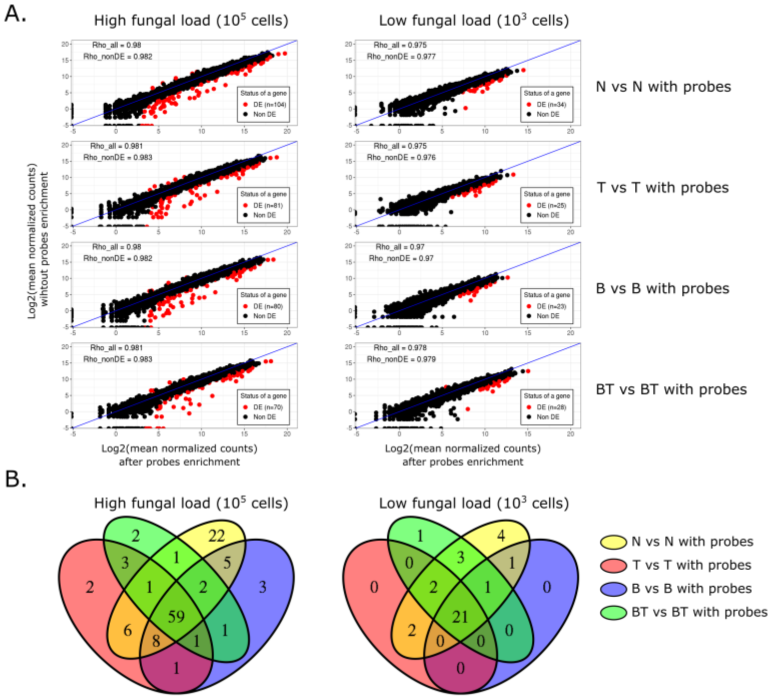


Fig. 6.3. (A) Correlation plots of average normalized fungal read counts between original and probe-enriched samples. DE: Differentially Expressed; Rho_all: Spearman's correlation coefficients calculated with all genes, Rho_nonDE: Spearman's correlations coefficients calculated with all genes, excluding differentially expressed ones; **(B)** Venn diagrams of DE genes between non-probe enriched and probe-enriched samples, indicating the effect of probe enrichment on expression levels of fungal genes.

As expected, treatment with the transcriptional inhibitor thiolutin did not affect the expression levels of genes, except for 1 to 13, which might be attributed to experimental manipulations.

Probe-enriched RNA-Seq data allow variant calling analysis

RNA-Seq data can be used to call variants in the transcriptionally active parts of the genome. Hence, we further tested whether it is feasible to perform variant calling analysis using the probe-enriched data. For this, we performed SNP calling analysis for all datasets and compared the results between the probe-enriched and non-enriched samples (suppl. Fig. S6.4). We observed that 50-75% of SNPs are retained after the probe enrichment. Moreover, 80-90% SNPs that were identified after probe enrichment were identical to the ones obtained from the non-enriched samples.

We further tested whether the identified variants in our data are sufficient to determine the genetic background of the cells that were spiked in the samples. To this end, we compared the variants identified in our samples with those present in a diverse set of 58 *C. albicans* strains (including SC5314), representing the major clades of its inter-strain phylogenetic tree (Mixão and Gabaldón, 2020).

As a measure of genetic relatedness, we calculated the number of shared variants between our sample and any given strain, relative to the total number of variants in the latter. The results (Fig. 6.4) of this analysis showed that the called variants before and after target enrichment provided enough resolution to identify the phylogenetic clades to which the strain belonged to (clade 1), although not the specific strain (Ropars *et al.*, 2018).

It must be noted that we observed similar results when using both enriched and non-enriched data, indicating that clade level identification, rather than identification of the specific strain, is due to a relatively limited resolution of RNA-Seq-based genotyping, and not because of probe-based enrichment.

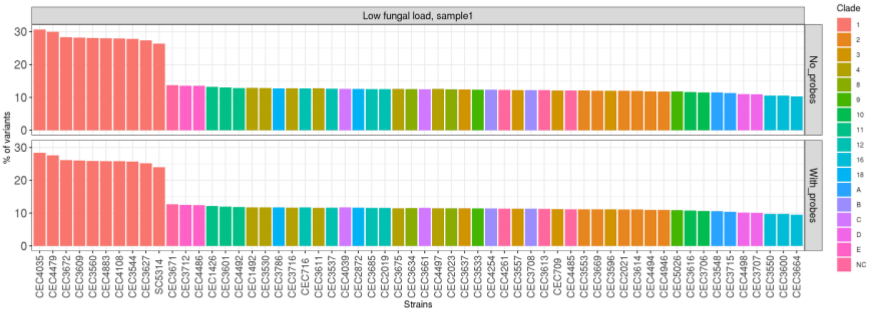


Fig. 6.4. Comparison of variants of enriched and non-enriched samples (treatment "N") with previously published *C. albicans* strains. Each bar represents the proportion obtained by dividing common variants between our sample and the published strain by the total number of variants of the published strain. Only the

data of one sample from the low fungal load are shown (see suppl. Fig. S6.5 for all samples from treatment "N").

Probe enrichment preserves qualitative patterns of microbiome distribution of analyzed samples

Considering that in our study the human samples were obtained from vaginal swabs, we set out to test whether our data after probe-based enrichment can be used for other types of analyses, such as for meta-transcriptome characterization. To test this, we used the RNA-Seq reads that were neither mapped to human nor to the *C. albicans* genome, to characterize the microbiome composition of probe-enriched and corresponding non-enriched samples.

For this, we mapped those reads to a custom database consisting of the human genome and the genomes of all bacteria, archaea and fungi, using Centrifuge, and calculated the relative abundance of identified genera in all samples. As a control, we performed a 16S rRNA-gene metagenomic analysis of one sample from the pool of vaginal swabs (Fig. 6.5B). The initial meta-transcriptomic analysis (suppl. Fig. S6.6) showed that abundances were dominated by reads for *H. sapiens* and *Candida dubliniensis,* a species closely related to *C. albicans*, whose genome is not present in the Centrifuge database. After filtering out the data of these two species, the meta-transcriptome analysis showed that, despite quantitative differences in abundance distribution, identified genera are qualitatively similar between enriched and non-enriched samples, and across lytic enrichments (Fig. 6.5A).

Moreover, after probe-based enrichment we identified bacterial general which are commonly found in vaginal microbiota. such as *Lactobacillus, Gardnerella* and *Prevotella,* among others, Furthermore, these results were confirmed with 16S rRNA-gene metagenomic analysis, which identified *Lactobacillus* and *Gardnerella* in the analyzed sample.

Overall, our results indicate that probe enrichment largely preserves the qualitative microbiome composition of analyzed samples, which can be used as a proxy for studying the microbiome as an additional layer of relevant information in the context of host-microbe interactions.

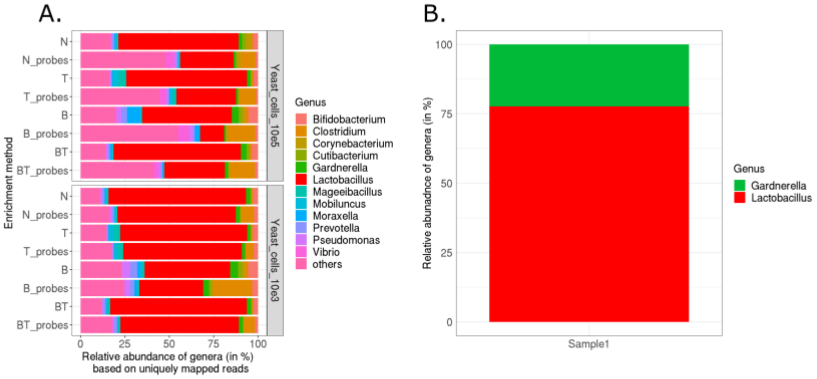


Fig. 6.5. Relative abundances of different genera identified in studied samples. **(A)** Results of meta-transcriptomic analysis. Each horizontal bar represents mean abundances across samples of a given enrichment type. Top 3 most abundant genera are considered for each sample. Row names indicate the enrichment type of the samples. Facet headers indicate fungal loads. **(B)** Results of 16S rRNA gene sequencing metagenomic analysis.

6.5 Discussion

Human-*Candida* interaction studies *in-vivo* are limited by the low proportion of fungal cells, a problem that is particularly important for transcriptome sequencing studies. Here, we addressed this issue by designing a probe-based enrichment targeting coding and non-coding transcriptomes of four main *Candida* pathogens. We tested our design using a large-scale dual RNA-Seq experimental setup, allowing us to compare the efficiency of the probe-based enrichment with another enrichment method, based on human cell lysis. We showed that in the context of enrichment efficiency, probe-based enrichment significantly outperforms the differential lysis approach. In addition, it retained the relative abundance of transcripts, biasing the expression levels of only ~0.3-1.5% of the genes. Thus, this approach can be reliably used for standard downstream transcriptomics analyses such as the detection of differential gene expression. As for differential human cell lysis approaches, we observed negligible enrichment achieved by these techniques. One possible explanation of this result could be that the initial testing of these methods was based on measuring the fungal ITS1 transcript, which in this study is removed by poly-A selection during the sequencing library preparation steps. This notion is confirmed by the fact that we did not detect any read mapping to ITS1 in any of our samples.

To date, there is only one study which also used targeted probe-capturing for pathogenic yeast transcriptome enrichment (particularly for *C. albicans*) to study host-pathogen interactions *in-vivo* by RNA-Seq. Amorim-Vaz *et al.* (2015) targeted the ORFome of *C. albicans* using the probe-based SureSelect technology (Agilent), which uses non-overlapping oligonucleotide probes, while SeqCap (Roche) uses overlapping probes for each target sequence. First the authors validated the probes, showing that probes alter expression levels of ~3% of *C. albicans* genes. Furthermore, probes were used to enrich fungal RNA from mouse kidney and *Galleria mellonella* infection models, demonstrating on average ~1130 fold enrichment of fungal reads. Although our fold enrichment is significantly lower, this is likely to be related to the higher initial relative amount of fungal RNA in the study of Amorim-Vaz *et al.* (2015), as we found a relationship between this quantity and the resulting fold enrichment. Taking this into consideration, we show that both approaches have comparable efficiencies.

We show that our design has several clear advantages. Firstly, our pan-*Candida* enrichment design not only targets the *C. albicans* transcriptome, but also has probes to capture transcriptomes of three other major pathogens *C. glabrata*, *C. parapsilosis* and *C. tropicalis*. Secondly, our design includes the newly predicted lncRNAs of these four species, opening new avenues for studying the potential role of these enigmatic molecules in host-pathogen interactions between the human host and *Candida* yeasts. Apart from thorough transcriptomic analysis, we asked whether the RNA-Seq data obtained after probe-based enrichment can serve for other types of analyses. Considering that RNA-Seq bears genotypic information of the transcribed regions of a genome, we assessed the feasibility of performing SNP calling analysis on our enriched data. Comparison between the pairs of non-enriched and enriched samples demonstrated that probe enrichment preserves 50-75% of original variants. Moreover, by comparing our genotyping data with those of other *C. albicans* strains, we show that it is possible to identify the infecting strain at a clade level. These results indicate that probe-enriched samples might additionally identify genotypic variants of the studied pathogen, which can serve as an important layer of information for fungal strain identification and antifungal susceptibility profiling.

Considering that we used vaginal samples for assessing the enrichment strategies, we tested whether probe-enriched data additionally serve to characterize microbial communities through meta-transcriptomics. The results of this analysis indicated that all types of tested enrichments largely retained the major microbial genera observed in the initial samples, such as *Lactobacillus* and *Gardnerella*. This was further confirmed by independent 16S rRNA metagenomic analysis. It has to be mentioned though that identification of microbial data after probe enrichment might be a side effect of probe capturing rather than intended result. The presence of microbial genetic material different from the one which was targeted by probes can be explained by a combination of several effects, such as non-specific probe binding or pull-down of random sequencing libraries along with the targeted ones. These unintended effects notwithstanding, the meta-transcriptomic results appeared to be systematic and consistent across different replicates and conditions and coincided with results of 16S rRNA profile. Overall, these results suggest that RNA-Seq data undergone probe enrichment can be used for qualitative assessment of microbial communities of the host.

6.6 Conclusions

We designed a pan-*Candida* probe-based enrichment approach based on SeqCap technology, targeting the coding and non-coding transcriptomes of the four major *Candida* pathogens. We extensively tested the kit using large-scale dual RNA-Seq of human vaginal samples spiked with *C. albicans* cells, showing its superior performance over the human cell lysis enrichment approach. Moreover, we demonstrated that RNA-Seq data generated after probe enrichment can serve as a source for additional valuable analyses such as fungal genotyping and host microbiome assessment. Our work highlights the power of targeted probe enrichment, which opens new horizons for investigating *in vivo* host-microbe interactions between human and *Candida* pathogens.

Acknowledgements

We thank Veronica Mixão for providing the variant calling data of published *C. albicans* strains. TG group acknowledges support from the Spanish Ministry of Science and Innovation for grants ‘Centro de Excelencia Severo Ochoa’ and PGC2018-099921-B-I00, cofounded by European Regional Development Fund (ERDF); from the CERCA Programme / Generalitat de Catalunya; from the Catalan Research Agency (AGAUR) SGR423, and grants from the European Union’s Horizon 2020 research and innovation programme under the grant agreement ERC-2016-724173, and the Marie Sklodowska-Curie grant agreement No H2020-MSCA-IF-2017-793699. The group also receives support from a INB Grant (PT17/0009/0023 - ISCIII-SGEFI/ERDF).

6.7 Supplementary figures

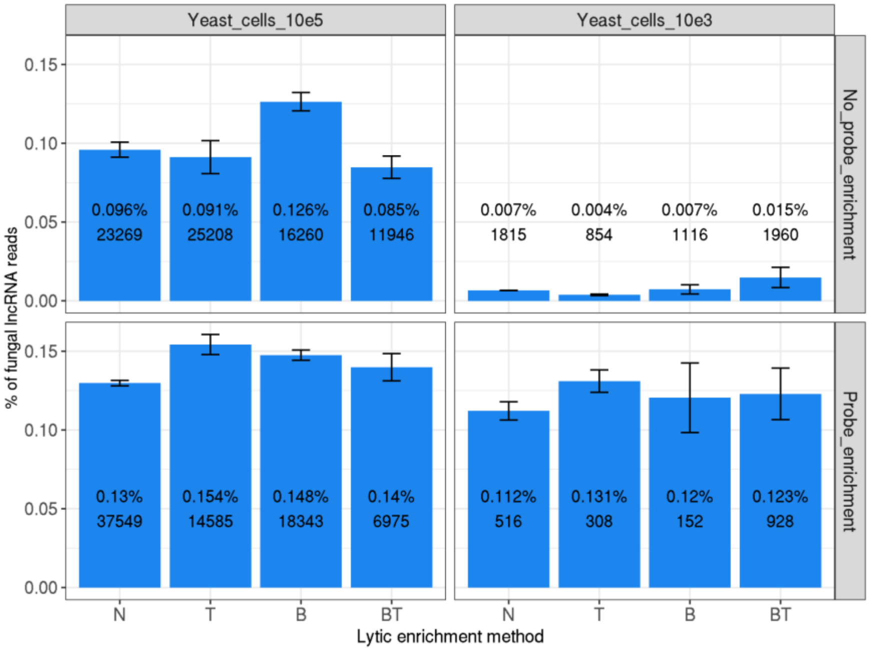


Fig. S6.1. The proportions of fungal lncRNA reads obtained with different enrichment approaches, as determined by RNA-Seq.

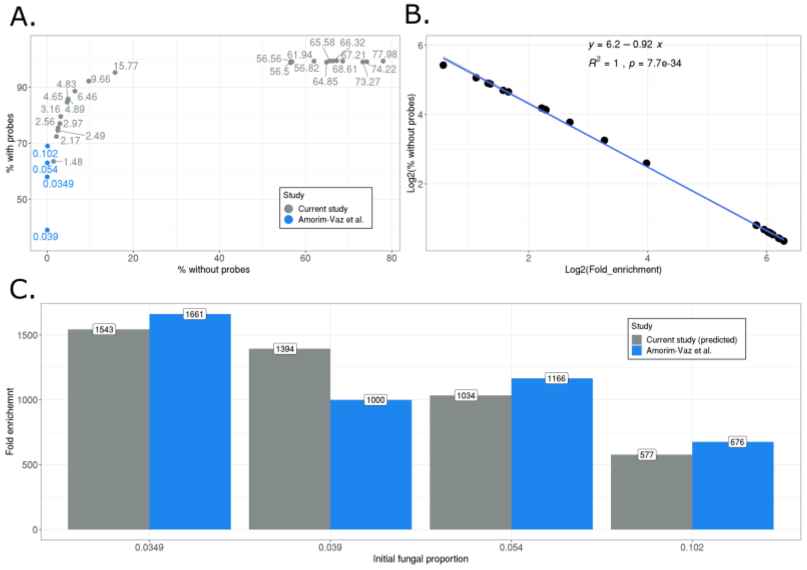


Fig. S6.2: Enrichment efficiency of our probe design compared to that of Amorim-Vaz *et al.* (2015). **(A)** Initial (i.e. before targeted enrichment) and final (i.e. after targeted enrichment) fungal proportions of the samples in both studies. Labels

indicate the initial fungal proportions; **(B)** Linear model of fold enrichment depending on the initial fungal proportions; **(C)** Comparison of fold enrichments observed in Amorim-Vaz *et al.* (2015), and predicted fold enrichment of our probes based on the linear model and initial fungal proportions reported in Amorim-Vaz *et al.* (2015).

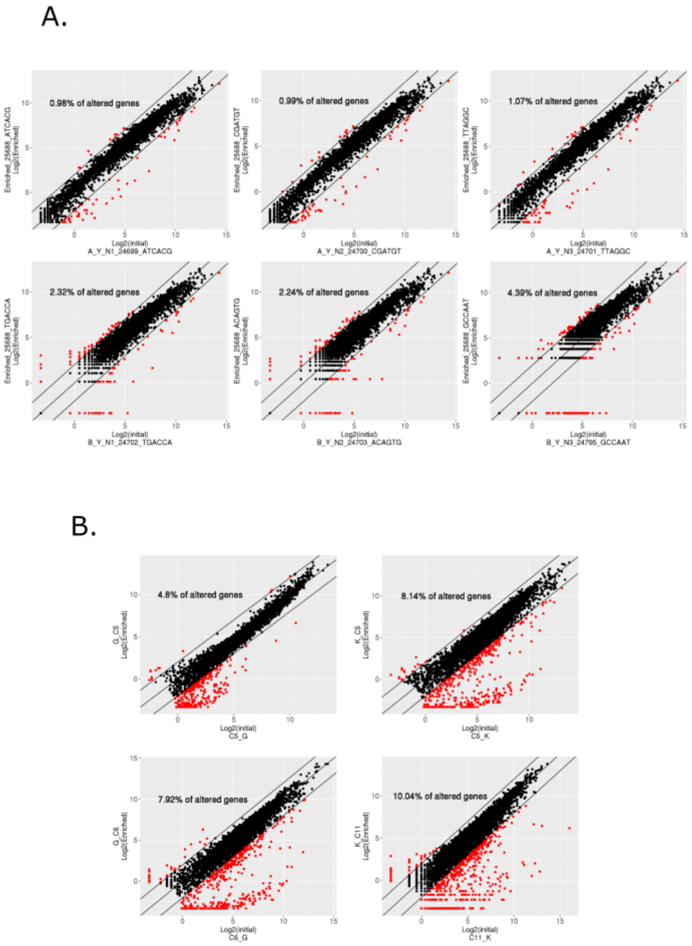


Fig. S6.3: Scatter plots displaying log2 normalized counts before and after probe-based enrichment. **(A)** Data of the current study (treatments “N”), top row: samples with high fungal load, bottom row: low fungal load, labels of the samples on axes: internal sample identifiers; **(B)** Data from Amorim-Vaz *et al.* (2015), labels on the axes: sample labels from that study.

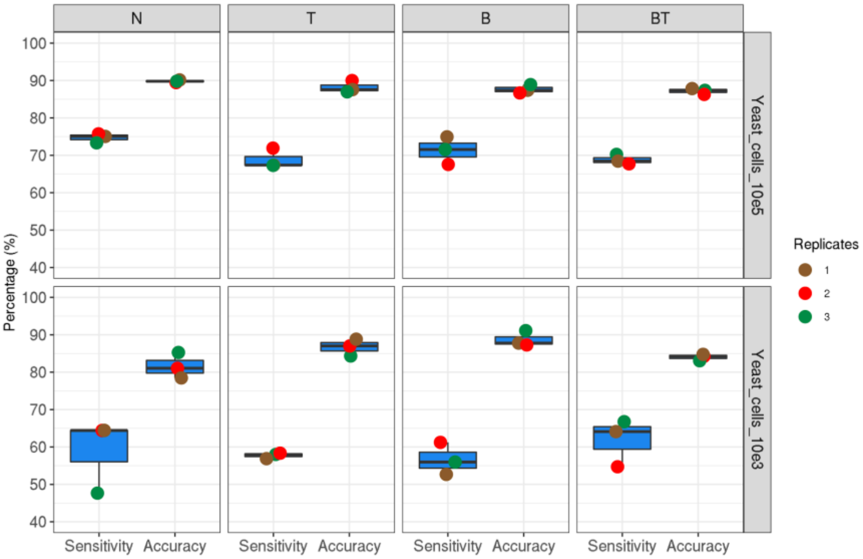


Fig. S6.4: Distribution of SNPs after probe-based enrichment compared to the non-enriched samples. The analysis is done between enriched and corresponding non-enriched samples of each replicate. The comparison between samples of each replicate is shown by a dot. Colors of dots correspond to the replicate number. Box plots show the distribution across replicates. Sensitivity: proportion of the number of common variants between enriched and non-enriched samples divided by the number of variants in the non-enriched sample. Accuracy: proportion of the number of common variants between enriched and non-enriched samples divided by the number of variants in the probe enriched sample.

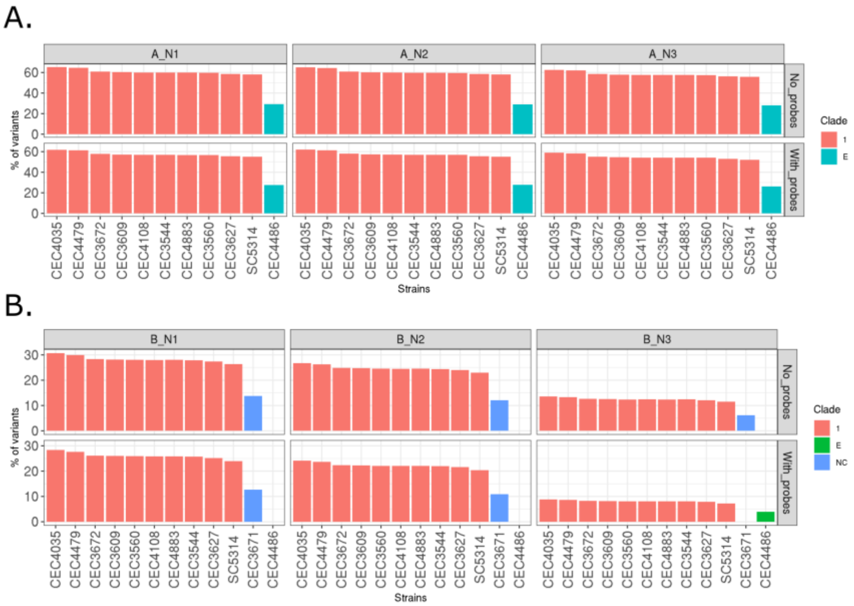


Fig. S6.5. Comparison of variants of enriched and non-enriched samples (treatment "N") with sequenced *C. albicans* strains (Ropars *et al.*, 2018). Each bar represents the proportion obtained by dividing common variants between our sample and the published strain by the total number of variants of the published strain. Only strains with top 11 proportions are plotted. **(A)** Results for high fungal load; **(B)** Results for low fungal load.

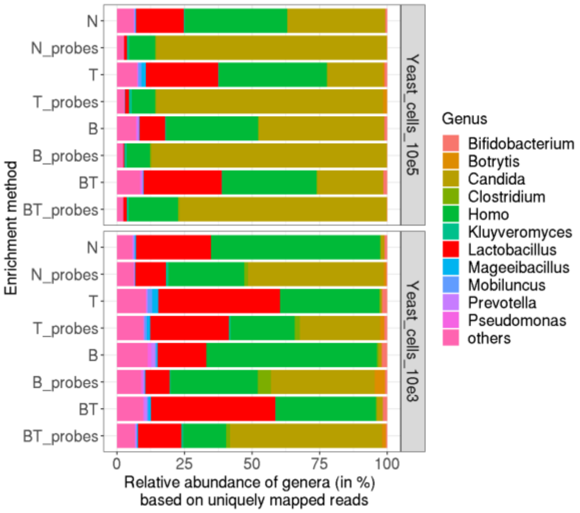


Fig. S6.6. The initial meta-transcriptomic analysis, demonstrating that meta-transcriptomic profiles of samples are dominated by *Homo* and *Candida dubliniensis*. The reference genome of *C. albicans* is not available in Centrifuge database.

7 The transcriptional aftermath in two independently formed hybrids of the opportunistic pathogen *Candida orthopsilosis*

Hovhannisyan, Hrant, Ester Saus, Ewa Ksiezopolska, and Toni Gabaldón. "The transcriptional aftermath in two independently formed hybrids of the opportunistic pathogen *Candida orthopsilosis*" *mSphere* 5, no. 3 (2020). DOI: 10.1128/mSphere.00282-20

7.1 Abstract

Interspecific hybridization can drive evolutionary adaptation to novel environments. The *Saccharomycotina* clade of budding yeasts includes many hybrid lineages, and hybridization has been proposed as a source for new pathogenic species. *Candida orthopsilosis* is an emerging opportunistic pathogen for which most clinical isolates are hybrids, each derived from one of at least four independent crosses between the same two parental lineages. To gain insight on the transcriptomic aftermath of hybridization in these pathogens, we analyzed allele-specific gene expression in two independently formed hybrid strains, and in a homozygous strain representative of one parental lineage. Our results show that the effect of hybridization on overall gene expression is rather limited, affecting ~4% of the studied genes. However, we identified a larger effect in terms of imbalanced allelic expression, affecting ~9.5% of the heterozygous genes in the hybrids. This effect was larger in the hybrid with more extensive loss of heterozygosity, which may indicate a tendency to avoid loss of heterozygosity in these genes. Consistently, the number of shared genes with allele-specific expression in the two independently formed hybrids was higher than random expectation, suggesting selective retention. Some of the imbalanced genes have functions related to pathogenicity, including zinc transport and superoxide dismutase activities. While it remains unclear whether the observed imbalanced genes play a role in virulence, our results suggest that differences in allele-specific expression may add an additional layer of phenotypic plasticity to traits related to virulence in *C. orthopsilosis* hybrids.

7.2 Introduction

The incidence of human fungal infections has steadily increased during the past decade, leading to recognition of their relevance in global epidemiology (Oren and Paul, 2014). Numerous factors may underlie this growing prevalence including, among others, increased number of immunocompromised individuals (elderly people, neonates, HIV, patients, etc) (Pfaller and Diekema, 2007), emergence of drug resistance associated with extensive use of antimycotic agents (Pfaller *et al.*, 2009; Ksiezopolska and Gabaldón, 2018), globalization (Callaghan and Guest, 2015; Mixão and Gabaldón, 2018), and climate change (Garcia-Solache and Casadevall, 2010). The rising incidence of mycoses is also coupled to the identification of a larger number of etiological agents, including so called emergent pathogens that are increasingly identified in the clinics (Papon *et al.*, 2013; Gabaldón, Naranjo-Ortíz and Marcet-Houben, 2016). In this context, hybridization between different species or lineages has been identified at the origin of several emerging yeast pathogens (Pryszcz *et al.*, 2014, 2015; Wu *et al.*, 2015; Schröder, Martinez de San Vicente, *et al.*, 2016; Mixão and Gabaldón, 2018; Mixão *et al.*, 2019).

For some fungal hybrid pathogens the two parental species have been identified, as in the case of *Cryptococcus neoformans* × *Cryptococcus deneoformans* (Boekhout *et al.*, 2001), and *Cryptococcus neoformans* × *Cryptococcus gattii* (Bovers *et al.*, 2008). For others, one or both (or possibly numerous) parentals are still unknown. For example, for *Candida orthopsilosis* only one of the two putative parental lineages has been identified (Pryszcz *et al.*, 2014; Schröder, Martinez de San Vicente, *et al.*, 2016), which constitutes a minority (~7%) of the analyzed clinical strains. In the case of *Candida metapsilosis* or *Candida inconspicua* both parentals remain unknown (Pryszcz *et al.*, 2015; Mixão *et al.*, 2019), as all analyzed strains are hybrids. The fact that some parental species of hybrid fungal pathogens are never or rarely identified among clinical isolates suggests that some human pathogens have arisen from non-pathogenic parental organisms (Pryszcz *et al.*, 2014; Mixão and Gabaldón, 2018), or that the parental lineages have been out-competed by their more adapted pathogenic hybrid descendants (Jonge *et al.*, 2013; J. R. Depotter *et al.*, 2016; J. R. L. Depotter *et al.*, 2016). In either case, whether interactions between the two distinct sub-genomes contribute to emerging properties in the hybrid, such as increased ability to infect humans, is still poorly understood.

At the molecular level, hybridization results in a state often referred as a “genomic shock” (McClintock, 1984), in which two diverged genomes which have evolved independently for a certain time are now sharing the same cellular environment. This coexistence can lead to alterations at several levels, including the genome (Dutta *et al.*, 2017), the transcriptome (Cox *et al.*, 2014), or the proteome (Hu *et al.*, 2017), among others (Greaves, Gonzalez-Bayon, Wang, Zhu, Liu, Groszmann, James Peacock, *et al.*, 2015). Advances in next-generation sequencing have facilitated the study of the genomic aftermath of hybridization, which involves phenomena such as large-scale genome duplications or deletions, homeologous recombination, and gene conversion leading to loss of heterozygosity (Marcet-Houben and Gabaldón, 2015; Pryszcz *et al.*, 2015; Dutta *et al.*, 2017; Smukowski Heil *et al.*, 2017; Mixão and Gabaldón, 2018).

However, our understanding of the impact of hybridization at the transcriptomic level remains poorly characterized, with few studies performed on industrial or plant saprophyte hybrids (Tirosh *et al.*, 2009; Cox *et al.*, 2014; Li and Fay, 2017; Metzger, Wittkopp and Coolon, 2017; Hovhannisyan, Saus, *et al.*, 2020).

A powerful approach to assess the impact of hybridization on the transcriptome is the assessment of allele-specific expression (ASE) - a phenomenon in which one of the two alleles of the gene is preferentially expressed over the other one. With current advances in next-generation sequencing technologies, such as RNA-Seq, ASE can be assessed in a transcriptome-wide manner. Numerous studies of ASE have been performed in recent years, including the ones in yeasts (Muzzey, Sherlock and Weissman, 2014; Salinas *et al.*, 2016; Thompson and Cubillos, 2017).

To date no study of the ASE and transcriptomic aftermath of hybridization has been performed in hybrid human pathogens, limiting our insights on how hybridization leads to emergent traits, including virulence. To fill in this gap, we here undertook a transcriptomic analysis of two independently formed hybrid strains from the emerging yeast pathogen *Candida orthopsilosis* (Pryszcz *et al.*, 2014; Schröder, Martinez de San Vicente, *et al.*, 2016). This yeast belongs to the CTG clade and is phylogenetically placed within the *C. parapsilosis sensu lato* species complex alongside with the other opportunistic pathogens *C. parapsilosis* (Pammi *et al.*, 2013; Tóth *et al.*, 2019) and *C. metapsilosis* (Tavanti *et al.*, 2005). It has been shown that most (~93%) of the clinical isolates of *C. orthopsilosis* are hybrids between parentals with ~5.1% nucleotide divergence. As mentioned above, only one of the two parental lineages has been found among clinical isolates (Schröder, Martinez de San Vicente, *et al.*, 2016). To date, the other partner in the hybridization remains unidentified. Notably, these two parental lineages have hybridized several independent times, giving rise to at least four distinct hybrid clades that differ in their levels and patterns of loss of heterozygosis (LOH) (Schröder, Martinez de San Vicente, *et al.*, 2016). Clade 1 comprises strains which underwent extensive LOH, and are thought to derive from a relatively ancient hybridization, while strains in clade 4 are the most heterozygous, with fewer LOH events and are thus assumed to result from a more recent event. Thus, *C. orthopsilosis* represents an appropriate model to study how hybridization impacts transcription in a natural hybrid pathogen, and whether parallel hybridization events result in similar transcriptomic interactions between the two parental sub-genomes.

To shed light on these questions, we conducted ASE analysis using RNA-Seq of two hybrid strains of *C. orthopsilosis* each resulting from an independent hybridization event (Fig. 7.1), that represent the two extremes of LOH extent, namely MCO456 (clade 1) and CP124 (clade 4). For comparison, we investigated the transcriptome of a highly homozygous strain belonging to one of the putative parental lineages (strain 90-125). To our knowledge this is the first description of the transcriptomic profiles between parental species in a hybrid yeast that is an opportunistic pathogen of humans.

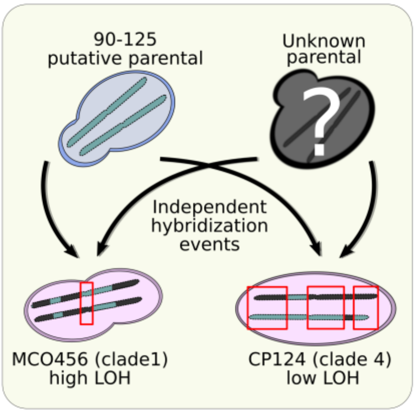


Fig. 7.1: Schematic representation of the experimental design of the study. The C. orthopsilosis 90-125 strain represents a putative parental lineage, which has undergone several independent hybridization events (black arrows) by mating with a second, unknown parental strain. A supposedly more ancient hybridization event has given rise to a hybrid clade, including the MCO456 strain ("high LOH"), which experienced extensive LOH, and a more recent hybridization event has led to the formation of an independent hybrid clade, including CP124 strain ("low LOH"), which contains more heterozygous regions (highlighted with red rectangles).

7.3 Results

To understand the impact of hybridization on gene expression, and disentangle the transcriptomic interactions between the two parental subgenomes in the pathogenic hybrid yeast *C. orthopsilosis*, we performed RNA-Seq of two hybrid strains, and a homozygous strain belonging to one of the putative parental lineages (Pryszcz *et al.*, 2014; Schröder, Martinez de San Vicente, *et al.*, 2016). Importantly, the two analyzed hybrid strains belong to two independently formed hybrid clades resulting from the mating of the same two parental lineages: strain MCO456 belongs to the hybrid clade 1, which underwent extensive LOH, whereas strain CP124 belongs to clade 4, which has limited LOH (Fig. 7.1). Additionally, we analyzed a strain belonging to one of the putative parental homozygous lineages (90-125) and compared its transcriptomic profile with the hybrids. The summary statistics of our RNA-Seq data including mapping rates and reproducibility metrics are available in suppl. Data Set S1, Tab 1 and Tab 2, respectively.

To perform accurate assignment of RNA-Seq reads to each of the two parental subgenomes in the hybrid, we used the following genome phasing procedure (see Fig. 7.2 and Materials and Methods for details). Knowing the genome of a relative of one of the putative parentals (i.e. strain 90-125), we used it as a reference to map publicly available DNA-Seq raw data of the two hybrid strains. We phased the heterozygous regions in each of the hybrids by reconstructing the haplotype belonging to the known parental lineage (i.e. having the alleles present in the 90-125 sequence) and the alternative haplotype belonging to the unknown parental lineage (i.e having the heterozygous alleles alternative to 90-125). Using this procedure, we could phase 107 and 590 genes within heterozygous regions for MCO456 and CP124 strains, respectively (suppl. Data Set S1, Tab 3 and Tab 4), from which 71 genes were common between the two strains. As expected, we obtained more phased genes in the clade 4 strain (CP124), which encompasses more heterozygous regions than MCO456 (Clade 1). Notably, the overlap of 71 genes in heterozygous regions between the two strains is more than expected by chance as calculated by hypergeometric test (p=1.011097e-45), which suggests the existence of structural or selective constraints acting on LOH events.

Once we obtained the phasing information, we further assessed the levels of allelic expression in hybrids by mapping the RNA-Seq data to the concatenated phased genomes using strict filters for read mapping mismatches (see Materials and Methods). We checked the rates of cross-mapping - i.e. reads that cannot be unambiguously mapped to either parental - by mapping the data from the parental strain to the concatenated phased genomes, and observed a negligible proportion of cross-mapping in phased genes (~0.019% and ~0.023% for MCO456 and CP124, respectively, suppl. Fig. S1 and S2).

The expression levels in 90-125 strain were obtained by mapping RNA-Seq data directly to the reference genome. Read counts for all strains can be found in suppl. Date Set S1, Tab 5 and Tab 6. Then we performed differential expression (DE) and allele-specific expression (ASE) analysis in both hybrid strains and the parental strain (Fig. 7.3).

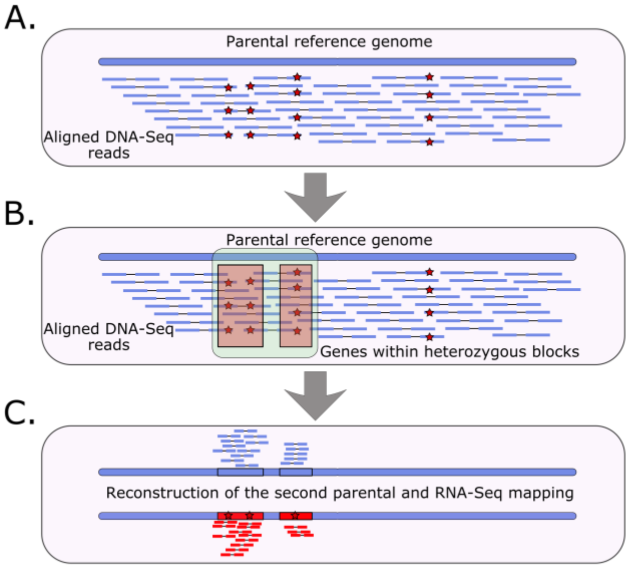

A.
Parental reference genome
Aligned DNA-Seq reads
B.
Parental reference genome
Aligned DNA-Seq reads
Genes within heterozygous blocks
C.
Reconstruction of the second parental and RNA-Seq mapping

Fig. 7.2. Schematic representation of the bioinformatics approach to assess the allele-specific expression in the hybrid strains. **(A)** Mapping of DNA-Seq reads to the parental reference genome and further variant calling (red stars represent heterozygous variants). **(B)** Defining heterozygous blocks (green rectangle) and identifying genes (red rectangles) within the blocks. **(C)** Inserting the heterozygous variants in the reference genome (second parental reconstruction) and further RNA-Seq read mapping to the partially phased genome.

We first compared the expression levels of 90-125 genes with the corresponding genes in heterozygous blocks in the hybrid strains. Such parental-hybrid comparisons (Fig. 7.3A and B) revealed a limited effect of hybridization on the parental gene expression levels - 5 (4.6%) and 18 (3%) DE genes when comparing 90-125 with MCO456 and CP124 hybrid, respectively (suppl. Data Set S1, Tab 7 and Tab 8).

Most of the differentially expressed genes have unknown roles, with some exceptions (suppl. Data Set S1, Tab 9 and Tab 10). For example, CORT_0B00460 which is up-regulated in the hybrid MCO456 background, is predicted to have functions related to metal ion binding and superoxide metabolic processes.

Further, we performed allele-specific gene expression analysis in the hybrids. We detected 11 allele-specifically expressed (ASE) genes in MCO456 (10.2% of phased genes, suppl. Data Set S1, Tab 11), while CP124 showed 51 ASE genes (8.6% of phased genes, suppl. Data Set S1, Tab 12).

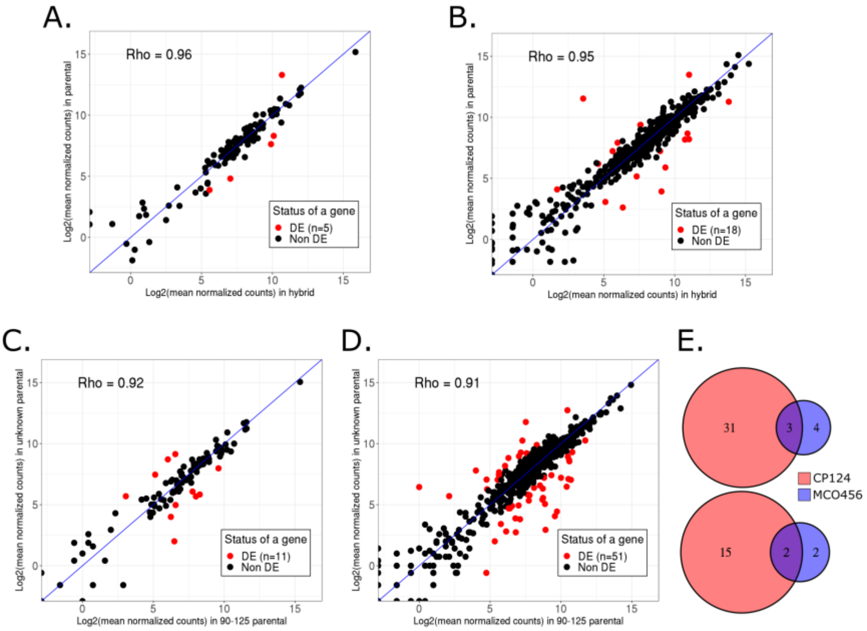


Fig. 7.3. Overall results of DE and ASE comparisons. (**A**) Correlation between gene expression levels in 90-125 and MCO456 (**B)** Correlation between gene expression levels in 90-125 and CP124. **(C)** ASE analysis for *C. orthopsilosis* MCO456 strain **(D)** ASE analysis for CP124 strain. Scatter plots are based on mean normalized read counts for each gene. "DE" denotes **D**ifferentially **E**xpressed. **(E)** Venn diagrams showing the overlap between ASE genes in both strains: up-regulated 90-125 homeologs (on the top) and up-regulated homeologs of unknown parent (at the bottom)

Putative functions for ASE genes in MCO456 and CP124 strains can be found in suppl. Data Set S1, Tab 13 and Tab 14, respectively. Although the function of most ASE genes was unknown, some have orthologs involved in virulence in other *Candida* pathogens. For both strains, genes related to superoxide dismutase activity (Martchenko *et al.*, 2004; Li *et al.*, 2015; Broxton and Culotta, 2016) (CORT_0B00460 in MCO456, and CORT_0A12390 in CP124 strain) were up-regulated for the allele of the unknown parent (Martchenko *et al.*, 2004; Li *et al.*, 2015; Broxton and Culotta, 2016). When comparing the 90-125 parental with the corresponding homeologs in the hybrids, the expression level of CORT_0A12390 was intact upon hybridization, while CORT_0B00460 was up-regulated in MCO456.

Moreover, we identified ASE genes related to zinc metabolism, which is one important player in host-pathogen interactions (Wilson, Citiulo and Hube, 2012; Crawford and Wilson, 2015; Jung, 2015). The gene CORT_0C02470, potentially involved in zinc transmembrane transport, was up-regulated in the hybrid towards the 90-125 parental allele in

MCO456, while CORT_0E02010, with a putative zinc binding activity, is expressed at higher levels from the allele assigned to the unknown parental in CP124. For both genes, the expression levels were not altered upon hybridization.

Notably, five genes - CORT_0A04580, CORT_0A09590, CORT_0A03280, CORT_0D04190, and CORT_0E01400, are common ASE genes between the two hybrids, which is higher than expected by chance considering the shared fraction of heterozygous genes (Fig. 7.3E, p=9.433106e-06). While the functions of the first four genes are unknown, the orthologs of CORT_0E01400 gene, which expresses more the unknown parental allele, are involved in cellular response to drugs and extracellular region localization (according to CGD annotations).

7.4 Discussion

Here, we investigated the transcriptomic interactions of divergent parental sub-genomes in *C. orthopsilosis* hybrids. To our knowledge this represents the first such study in a hybrid human opportunistic yeast pathogen. Our experimental design allowed us not only to compare expression of genes in hybrid genetic background to that in homozygous parental background, but also to assess the extent of convergent evolution in two independently formed hybrid lineages.

In agreement with previous studies of transcriptomic shock in fungal hybrids (Cox *et al.*, 2014; Hovhannisyan, Saus, *et al.*, 2020), our results indicate that hybridization has a rather moderate effect on gene expression levels – on average ~4% of the studied genes changed the expression levels upon hybridization, as compared to the parental. This relatively low levels of transcriptomic alteration upon hybridization in yeast hybrids, contrasts with the larger levels reported in animals or plants, as noted earlier (Hovhannisyan, Saus, *et al.*, 2020), and reinforcing the idea that yeasts have a comparatively higher capacity than plants and animals to buffer the effects of the transcriptomic shock elicited by hybridization. As more transcriptomic studies on diverse organisms accumulate, it will be clarified how widespread these differences are and what molecular mechanisms may underlie this phenomenon.

When comparing the results of this study with those of the report assessing transcriptomic shock in artificial yeast hybrid (Hovhannisyan, Saus, *et al.*, 2020), we noted that the proportion of DE genes (calculated over the total number of genes in heterozygous regions) in natural hybrids is somewhat larger (~4% in this study), than in the case of the artificial *S. cerevisiae x S.*

uvarum hybrids (~1.5%), despite the larger parental divergence and more recent nature of the hybridization. Additionally, we found that the two independently formed *C. orthopsilosis* hybrids, retained a shared subset of genes in heterozygosis, which is larger than expected by chance. Altogether these results suggest that structural constraints or functional selection may play a role in shaping LOH patterns in these hybrids, with certain genes, including some showing divergent expression patterns, being more likely to be retained in heterozygosis. Moreover, it has been previously reported that LOH events can be driven by selective pressures (Smukowski Heil *et al.*, 2017). A plausible scenario is that shared genes in heterozygosis, and shared DE genes between the two independently formed hybrids are involved in traits beneficial for the hybrid, and are thus maintained through purifying selection. Nevertheless, comparisons from such limited number of studies must be taken with caution, and we hope that future studies will help to clarify such questions.

When assessing differences in expression between homeologous copies in the hybrid (ASE), we found that the fraction of genes with ASE (8.6% in CP124, and 10.2% in MCO456) was comparatively larger than the fraction of DE genes. In addition, these fractions of ASE genes are slightly higher than that noted for a newly formed *S. cerevisiae x S. uvarum* hybrid (7.4%), despite the much lower parental divergence in *C. orthopsilosis* hybrids. Moreover, the fraction of ASE genes is larger in MCO456 (10.2% as compared to 8.6% in CP124) which underwent more extensive LOH. Although more datasets are needed to confirm this trend, this observation suggests that LOH preferentially targets genes that do not display ASE, consistent with the observation above for DE genes. Of note, we did not observe a strong preferential expression of alleles from either of parental subgenomes with 34 (~66%) and 7 (~63%) of ASE genes were expressed higher in the known parental in CP124 and MCO456 strains, respectively. This is in line with initial comparisons of *C. orthopsilosis* genomes, which showed no preferential retention of any of the parental genomes in regions undergoing LOH (Pryszcz *et al.*, 2014). Previous studies in natural (*Epichloë*) and artificial (*S. cerevisiae x S. uvarum*) fungal hybrids have also reported no strong preferential over-expression of one of the parental sub-genomes (Cox *et al.*, 2014; Li and Fay, 2017), which suggests this may be a general phenomenon in ascomycete fungal hybrids.

For both *C. orthopsilosis* hybrid strains we identified ASE genes involved in processes directly related to virulence, such as zinc ion transport and superoxide dismutase activity. However, since the second parental organism is still unidentified it is unknown whether the observed expression divergence between homeologs arose due to hybridization or whether it was already existing between orthologs of the parental species. One limitation of our study is that the conditions used do not fully resemble those of infection. However previous studies have reported that the transcriptional profile of *C. albicans* growing in YPD is remarkably similar to the one during interaction with the host during infection (Liu *et al.*, 2015). Thus, although it remains to be clarified whether the observed transcriptomic differences are actually related to differences in virulence, we argue that this possibility should be considered in future studies.

7.5 Conclusions

Altogether, our analyses provide evidence of a moderate effect of hybridization on the transcriptome of pathogenic hybrid yeasts, in line with previous observations in other fungal hybrids. Interestingly, the significant overlap of genes in heterozygous blocks, including DE and ASE genes, in the two independently formed hybrids suggests the existence of selecting constraints acting on genes that show altered expression in the hybrid. Similarly, we detected that increase in LOH was associated with higher fractions of ASE genes, suggesting that ASE genes are preferentially retained in heterozygosis in the hybrid. Finally, we detected ASE genes in the studied pathogen which are known to have direct implications in fungal virulence in other yeast species, like *C. albicans*, making them a target for further studies of fungal pathogenicity emergence.

7.6 Materials and Methods

Strains

We analyzed three diploid strains of *C. orthopsilosis* - MCO456, CP124, and 90-125, with the latter belonging to the lineage of one of the putative parentals of the two former hybrid species (Tavanti *et al.*, 2005; Riccombeni *et al.*, 2012; Pryszcz *et al.*, 2014).

Experimental conditions and RNA extraction

RNA extraction was performed on the samples growing at the exponential phase in rich yeast extract peptone dextrose medium (YPD) at 30ºC.
Experiments were performed as follows:
First, we measured growth curves for each individual strain to delimit the mid-exponential growth phase. For this, each strain was plated on a YPD agar plate and grown form 3 days at 30ºC. Single colonies were cultivated in 15 mL YPD medium in an orbital shaker (30ºC, 200 rpm, overnight). Then, each sample was diluted to an optical density at 600nm (OD) of 0.2 in 50 mL of YPD and then grown for 3h at the same conditions. Then

samples were diluted again to OD of 0.1 in 50 mL of YPD to start all experiments with a similar amount of cells. Cultures were grown at 30ºC for 24 hours and OD was every 60 min with a TECAN Infinite M200 microplate reader. Upon reaching the mid-exponential phase, the protocol was repeated until all samples were growing at the exponential phase. Then cultures were centrifuged at 16 000g to harvest 3x10$_8$ cells per sample. Total RNA from all samples was extracted using the RiboPure RNA Yeast Purification Kit (ThermoFisher Scientific) according to manufacturer's instructions. Total RNA integrity and quantity of the samples were assessed using the Agilent 2100 Bioanalyzer with the RNA 6000 Nano LabChip Kit (Agilent) and NanoDrop 1000 Spectrophotometer (Thermo Scientific).

RNA-Seq library preparation and sequencing

Sequencing libraries were prepared using the TruSeq Stranded mRNA Sample Prep Kit v2 (ref. RS-122-2101/2, Illumina) according to the manufacturer's instructions (unless specified otherwise). 1 µg of total RNA was used for poly(A)-mRNA selection using streptavidin-coated magnetic beads. Then samples were fragmented to approximately 300bp and cDNA was synthesized using reverse transcriptase (SuperScript II, Invitrogen) and random primers. The second strand of the cDNA incorporated dUTP in place of dTTP. Double-stranded DNA was further used for library preparation. dsDNA was subjected to A-tailing and ligation of the barcoded Truseq adapters. All purification steps were performed using AMPure XP Beads (Agencourt). Library amplification was performed by PCR on the size selected fragments using the primer cocktail supplied in the kit. To estimate the quantity and check the fragment size libraries were analyzed using Agilent DNA 1000 chip (Agilent), and were subsequently quantified by qPCR using the KAPA Library Quantification Kit (KapaBiosystems) prior to amplification with Illumina’s cBot. Libraries were loaded and sequenced using 2x50 or 2x75 read lengths on Illumina’s HiSeq 2500. Experiments were performed in three biological replicates. All library preparation and sequencing steps were performed at the Genomics Unit of the Centre for Genomic Regulation (CRG), Barcelona, Spain.

Bioinformatics analysis
Quality control of sequencing data

We used FastQC v0.11.6 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and Multiqc v. 1.0 (Ewels *et al.*, 2016) to perform quality control of raw sequencing data.

Phasing of heterozygous genomic regions

The reference genome and genome annotations for the putative parent strain *C. orthopsilosis* 90-125 were obtained from NCBI (assembly ASM31587v1, last accessed on 12/08/2018). To assess allele-specific expression in the hybrid strains we phased the hybrid genomes following the procedure illustrated in Fig. 7.2 and further described below.

Specifically, we first phased (i.e. resolved alternative haplotypes) the genes located in the heterozygous regions in the MCO456 and CP124 hybrid strains. To do this, we used DNA sequencing data of these strains (ERR321926 (Pryszcz *et al.*, 2014, 2015) and SRR3547561 (Schröder, Martinez de San Vicente, *et al.*, 2016)). We first trimmed these data using Trimmomatic v. 0.36 software with default parameters, and subsequently processed the data using mapping and variant calling modules of HaploTypo v.1 pipeline, setting the filter of SNP clusters at 5 SNPs in 20 bp window (Pegueroles, Mixão, *et al.*, 2019), and subsequently removed indels using vcftools v0.1.16 (Danecek *et al.*, 2011). Then using the heterozygous variants we defined heterozygous blocks using bedtools v2.29 (Quinlan and Hall, 2010) with *merge* function as described in (Pryszcz *et al.*, 2015) and further optimized in (Mixão *et al.*, 2019) – if the distance between two heterozygous variants is less than 100 base pairs (bp), that region constitutes a heterozygous block, if the next variant to that block is located in less than 100 bp, the block is extended, otherwise it is interrupted. After defining heterozygous blocks, we identified genes located within the blocks using a custom python script find_genes_in_heterozygous_blocks.py v1.0 (available at https://github.com/Gabaldonlab/C.-orthopsilosis-ASE/blob/master/find_genes_in_heterozygous_blocks.py). Samtools v. 1.9 was used to index the reference genome. Sorting of vcf files was done by *sort* function in bash. Subsequently, we inserted the alternative variants of the genes located in heterozygous blocks in the reference genome of the 90-125 strain using GATK v.3.7 (DePristo *et al.*, 2011), thus reconstructing the sequence of the alternative parental genome within the defined heterozygous blocks.

RNA-Seq and allele-specific expression analysis

RNA-Seq read mapping and summarization was performed using the splice-junction aware mapper STAR v. 2.7.3a (Dobin *et al.*, 2013). GFF to GTF format conversion for genome annotations was done using gffread v. 0.11.6 (Trapnell *et al.*, 2012) utility. For 90-125 strain, we mapped RNA-Seq data to the 90-125 reference assembly, while for the hybrid strains the data were mapped to a concatenated reference genome, obtained by combining the 90-125 reference and the reconstructed parental reference.

For mapping to the concatenated reference, we set STAR option --outFilterMismatchNmax to 0 to restrict the mismatches in read alignments. Differential gene expression and allele-specific expression were assessed using DESeq2 v. 1.22.2 (Love, Huber and Anders, 2014). For allele-specific expression comparisons the sizeFactors were set to 1 for all the samples, since the read counts for alleles come from the same library. For a gene to be considered differentially (allele-specifically) expressed, we used a threshold of |log2 fold change (L2FC)| > 1.5 and padj (adjusted p-value) < 0.01. Hypergeometric tests were performed using phyper function of R with lower.tail and log.p parameters set to FALSE.

To visualize gene expression data we used ggplot2 v. 2_3.0.0 R library (Wickham, 2016). Putative functions of *C. orthopsilosis* genes were retrieved from *Candida* Genome Database (Skrzypek *et al.*, 2017).

Data availability

RNA-Seq data is deposited at the SRA database under the accession numbers SRR10251160-SRR10251168.

Supplementary information can be found at https://msphere.asm.org/content/5/3/e00282-20/figures-only

Acknowledgements

We would like to thank Verónica Mixão (BSC/IRB, Barcelona, Spain) for her help in performing analysis with the HaploTypo pipeline. This work received funding from the European Union’s Horizon 2020 Research and Innovation Programme 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 Programme/Generalitat de Catalunya; from the Catalan Research Agency (AGAUR) SGR857, and grants from the European Union’s Horizon 2020 Research and Innovation Programme under the Grant Agreement No. ERC-2016-724173. TG also receives support from a INB grant (PT17/0009/0023 – ISCIII-SGEFI/ERDF).

8 Integrative omics analysis reveals a limited transcriptional shock after yeast interspecies hybridization

Hovhannisyan, Hrant, Ester Saus, Ewa Ksiezopolska, Alex J. Hinks Roberts, Edward J. Louis, and Toni Gabaldón. "Integrative omics analysis reveals a limited transcriptional shock after yeast interspecies hybridization." *Frontiers in Genetics* 11 (2020). DOI: 10.3389/fgene.2020.00404

8.1 Abstract

The formation of inter-specific hybrids results in the coexistence of two diverged genomes within the same nucleus. It has been hypothesized that negative epistatic interactions and regulatory interferences between the two sub-genomes may elicit a so-called “genomic shock” involving, among other alterations, broad transcriptional changes. To assess the magnitude of this shock in hybrid yeasts, we investigated transcriptomic differences between a newly formed *Saccharomyces cerevisiae x Saccharomyces uvarum* diploid hybrid and its diploid parentals, which diverged ~20 million years ago. RNA-Seq based allele-specific expression analysis indicated that gene expression changes in the hybrid genome are limited, with only ~1-2% of genes significantly altering their expression with respect to a non-hybrid context. In comparison, a thermal shock altered six times more genes. Furthermore, differences in expression between orthologous genes in the two parental species tended to be diminished for the corresponding homeologous genes in the hybrid. Finally, and consistent with RNA-Seq results, we show a limited impact of hybridization on chromatin accessibility patterns, as assessed with ATAC-Seq. Overall, our results suggest a limited genomic shock in newly formed yeast hybrid, which may explain the high frequency of successful hybridization in these organisms.

Key words: hybridization, yeast hybrid, transcriptome shock, allele-specific expression, buffering

8.2 Introduction

Inter-specific hybridization, meaning the mating of two different species to produce viable offspring, has been observed across a wide range of eukaryotic taxa, and is considered a major mechanism driving adaptation to new environmental niches (Gladieux *et al.*, 2014; J. R. Depotter *et al.*, 2016;

Session *et al.*, 2016). Hybridization in animals (Schwenk, Brede and Streit, 2008) and plants (Rieseberg, 1997) have long been recognized, and these organisms have focused the attention of most of the studies addressing the mechanisms and consequences of hybridization. In contrast, hybridization in microbial eukaryotes has been historically neglected, given the difficulty to detect morphological or physiological differences between species and their hybrids. It was the deep physiological and genetic characterization of the model yeast species *Saccharomyces cerevisiae* that allowed the discovery that several strains, initially classified as independent species, were in fact hybrids (Dujon, 2010). More recently, advances in next-generation sequencing (Goodwin, McPherson and McCombie, 2016) have facilitated the discovery of hybrids, demonstrating that hybridization is more frequent than previously anticipated, particularly in some microbial groups such as fungi (Albertin and Marullo, 2012). *Saccharomycotina* yeasts seem particularly prone to hybridization (Morales and Dujon, 2012), and there are numerous examples of yeast hybrid lineages of clinical (Pryszcz *et al.*, 2014, 2015; Schröder, Martinez de San Vicente, *et al.*, 2016; Mixão *et al.*, 2019) or industrial (Le Jeune *et al.*, 2007; Baker *et al.*, 2015; Krogerus *et al.*, 2017) relevance. Furthermore, a hybridization event has been proposed to have led to an ancient whole genome duplication in the lineage leading to *S. cerevisiae* and related yeasts (Marcet-Houben and Gabaldón, 2015).

An immediate outcome of inter-species hybridization is the coexistence of divergent genetic material within the same nucleus. This has been proposed to lead to a state called “genomic shock” (McClintock, 1984), in which negative epistatic interactions between the two coexisting sub-genomes, including interference between their gene regulatory networks, result in large physiological alterations.

Recent research has studied the effects of this “shock” on different layers of cellular organization, including, among others, the genome (Dutta *et al.*, 2017; Smukowski Heil *et al.*, 2017), the transcriptome (Cox *et al.*, 2014; Hu *et al.*, 2016; Lopez-Maestre *et al.*, 2017), the epigenome (Groszmann *et al.*, 2011; Greaves, Gonzalez-Bayon, Wang, Zhu, Liu, Groszmann, Peacock, *et al.*, 2015), and the proteome (Guo *et al.*, 2013; Hu *et al.*, 2017). Specifically, the assessment of transcriptomic changes in hybrids has been used for exploring *cis*- and *trans*- regulation of gene expression (Tirosh *et al.*, 2009; Graze *et al.*, 2012; Li and Fay, 2017; Metzger, Wittkopp and Coolon, 2017; Waters *et al.*, 2017). The comparison of gene expression levels in hybrid lineages versus their respective parents constitutes a versatile model for assessing gene regulation (Wittkopp, Haerum and Clark, 2004). Considering that parental genomes in a hybrid are exposed to the same cellular environment, and thus *trans*- regulatory elements, the difference in gene expression levels within a hybrid can be attributed to *cis*-regulation, while differences observed between parental organisms are due to a combination of *cis*- and *trans*- effects (Wittkopp, Haerum and Clark, 2004). Using this concept, *cis*- and *trans*- regulatory effects on gene expression have been studied in numerous taxa, including fungi (Thompson and Regev, 2009), flies (McManus *et al.*, 2010) and plants (Guo *et al.*, 2008; Combes *et al.*, 2015). Most transcriptomic studies of fungal hybrids have been performed in that particular context. For instance, Tirosh and colleagues (Tirosh *et al.*, 2009) investigated the impact of *cis*- and *trans*-effects on gene expression divergence in closely related *S. cerevisiae* and *S. paradoxus* and their interspecific hybrid at four different growth conditions. By performing within-hybrid (*cis*- effects) comparisons and subtracting those from between-parent comparisons (*trans*- effects), the authors demonstrated that the majority of regulatory divergence was the result of *cis*- effects, attributed to differences in promoter and regulatory regions that were independent of the environmental condition. On the other hand, *trans*- effects were related to transcription and chromatin regulators and were mostly condition-specific.

Using a similar approach, Metzger et al., (Metzger, Wittkopp and Coolon, 2017) used publicly available RNA-Seq datasets of two *S. cerevisiae* strains and their hybrid (Schaefke *et al.*, 2013) and data of *S. cerevisiae*, *S. paradoxus*, *S. mikatae* and *S. bayanus* and their respective hybrids (Schraiber *et al.*, 2013) to assess the dynamics of regulatory changes throughout long evolutionary distances. They concluded that as sequence divergence increases, *cis*- regulatory divergence becomes the dominant regulatory mechanism, and that both differences in gene expression and regulatory sequences increase with genetic distance, reaching a plateau for distantly related species.

Another study (Li and Fay, 2017) used *S. cerevisiae x S. uvarum* hybrid, resulting from the mating of two thermally divergent species, to investigate the effect of temperature on allele-specific expression (ASE). Using RNA-Seq, the authors assessed ASE patterns in the hybrids grown at different temperatures, and showed that most of the *cis*- divergence is temperature-independent, with only a small fraction of ASE genes influenced by thermal condition. Overall, most previous studies used transcriptomics of hybrids as a means to investigate *cis*- and *trans*- effects on gene regulation at various conditions and evolutionary distances, but they did not directly assess the impact of hybridization on gene expression and how this compares with the regulatory impact of other stresses. Given their different focus, these studies do not measure gene expression in matched parental pairs and their hybrids across different conditions, preventing the re-analysis of their data for the purpose of assessing the impact of hybridization and how it compares with environmental effects.

The direct consequences of hybridization on gene expression profiles of parental species have been mostly studied in plants and animals (McManus *et al.*, 2010; Yoo, Szadkowski and Wendel, 2013; A. Li *et al.*, 2014; Wu *et al.*, 2018; Zhang *et al.*, 2018). Though using different methodologies, all these studies report widespread transcriptomic changes following hybridization, 10-30% of the genes being significantly affected. In this context, fungal studies are more limited. Cox et al., (Cox *et al.*, 2014) did address this issue in the natural fungal diploid hybrid (allopolyploid) *Epichloë* Lp1 by comparing its expression patterns with those in its haploid parental species. The authors found that this natural hybrid retained most gene copies of the two parental species, and, most importantly, that these genes generally retained the gene expression levels from the parental counterparts. In addition, differences in expression between homeologous genes tended to be lower than the corresponding differences between the orthologous genes in the parental species. Based on these findings the authors concluded that the transcriptional response to hybridization was largely buffered. However, being based on a natural hybrid, this study does not allow to discard the possibility that the lack of strong differences in gene expression are due to amelioration through compensatory mutations subsequent to the hybridization. In addition, by comparing a diploid hybrid to haploid parentals that study could not disentangle the effects of ploidy change from those of hybridization.

We here set out to directly assess the immediate transcriptional impact of hybridization and compare it with the effect of an environmental stress. To this end, we conducted an integrative multi-omics study comparing two distantly related fungal species - *S. cerevisiae* (SC) and *S. uvarum* (SU) - and their newly made hybrid at two thermal conditions. Using RNA-Seq we assessed transcriptional differences between orthologous genes in the parental species, between genes in the parental and the hybrid genetic background, and between homeologous genes coexisting in the hybrid. To compare the relative impact of hybridization with an environmental stress, we performed these experiments in two different temperatures, of which one affects the two parental species differently. We further investigated the consequences of hybridization on chromatin states by performing an assay for transposase-accessible chromatin using sequencing (ATAC-Seq) and integrated its results with our RNA-Seq data to get mechanistic insights behind the transcriptomic alterations caused by inter-specific hybridization.

8.3 Materials and methods

Strains

The diploid hybrids of *S. cerevisiae* and *S. uvarum* were generated as follows: genetically tractable isogenic MATa and MATalpha haploids of the North American *S. cerevisiae* strain, YPS128, isolated from the bark of an oak tree, were previously generated (Cubillos, Louis and Liti, 2009; Liti *et al.*, 2009) by first isolating a single meiotic spore from the wild type homothallic strain resulting in complete homozygosity across the genome except for the MAT locus. The HO gene was then replaced by a Hygromycin resistance cassette resulting in a diploid heterozygous for HO. Haploid spores (ho::HYG MATa or MATalpha) were then isolated from this and URA3 was replaced in these by the G418 resistance cassette KANMX. Similarly the *S. uvarum* strain UWOPS99-807.1.1, isolated from Argentina, was dealt with in the same way resulting in isogenic haploids of both mating types (Wimalasena *et al.*, 2014). Diploid hybrids were formed by mating the MATa *S. cerevisiae* to the MATalpha *S. uvarum* and vice versa.

Experimental conditions and RNA extraction

Samples for RNA extraction were collected during mid-exponential growth phase in rich medium (YPD) at two different temperatures: 30 *ºC* (normal growing temperature for those species (Salvadó *et al.*, 2011)), and 12 *ºC* ("cold-shock" condition).

Experiments were performed as follows: first, to delimit the timing of mid-exponential growth phase, growth curves were obtained for each considered strain individually. For this, each strain was streaked from our glycerol stock collection onto a YPD agar plate and grown for 3 days at 30 *ºC*. Single colonies were cultivated in 15 mL YPD medium in an orbital shaker (30 *ºC*, 200 rpm, overnight). Then, each sample was diluted to an optical density of 600nm (OD) of 0.2 in 50 mL of YPD and grown for 3 h in the same conditions (30 *ºC*, 200 rpm). Then, dilutions were made again to reach an OD of 0.1 in 50 mL of YPD, in order to start all experiments with approximately the same amount of cells. The increasing growth was investigated in parallel with manual measurement of the OD from the 50 mL samples and in 100 µL samples by a microplate reader (TECAN Infinite M200). For manual measurements, we inspected the absorbance in 1 mL every 1h. For automated measurements, the samples were centrifuged for 2 min at 3000 g, washed with 1 mL of sterile water, and centrifuged again for 2 min at 3000 g. The pellet was resuspended in 1 mL of sterile water. Finally, 5 µL of each sample was inoculated in 95 µL of YPD in a 96-well plate. All experiments were run in triplicate. Cultures were grown in 96-well plates at 30 ºC for 24 hours and monitored to determine the OD every

10 min with the microplate reader. Both manual and automated OD readouts showed similar growth patterns.

Once the mid-exponential phase was determined at around 5 hours for all three species, the above-mentioned protocol was repeated until all samples were growing at the exponential phase and then the cultures were centrifuged at maximum speed of 16 000 g to harvest 3x10$_8$ cells per sample. For the cold-shock experiments, when samples reached the mid-exponential phase, they were grown for 5 h more at 12 *ºC*, and then the cells were harvested as described above. Total RNA from all samples was extracted using the RiboPure RNA Yeast Purification Kit (ThermoFisher Scientific) according to manufacturer's instructions. Total RNA integrity and quantity of the samples were assessed using the Agilent 2100 Bioanalyzer with the RNA 6000 Nano LabChip Kit (Agilent) and NanoDrop 1000 Spectrophotometer (Thermo Scientific).

RNA-Seq library preparation and sequencing

Libraries were prepared using the TruSeq Stranded mRNA Sample Prep Kit v2 (ref. RS-122-2101/2, Illumina) according to the manufacturer's protocol. All reagents subsequently mentioned are from this kit unless specified otherwise. 1 µg of total RNA was used for poly(A)-mRNA selection using streptavidin-coated magnetic beads. Subsequently, samples were fragmented to approximately 300bp. cDNA was synthesized using reverse transcriptase (SuperScript II, Invitrogen) and random primers. The second strand of the cDNA incorporated dUTP in place of dTTP. Double-stranded DNA was further used for library preparation. dsDNA was subjected to A-tailing and ligation of the barcoded Truseq adapters. All purification steps were performed using AMPure XP Beads (Agencourt). Library amplification was performed by PCR on the size selected fragments using the primer cocktail supplied in the kit. Final libraries were analyzed using Agilent DNA 1000 chip (Agilent) to estimate the quantity and check fragment size distribution, and were then quantified by qPCR using the KAPA Library Quantification Kit (KapaBiosystems) prior to amplification with Illumina's cBot. Libraries were loaded and sequenced 2x50 or 2x75 on Illumina's HiSeq 2500.

Genome-wide chromatin accessibility profiling by ATAC-Seq

The two studied strains, namely SC and the hybrid, were grown to mid-exponential phase in YPD at 30 *ºC* as described above and subjected to an Assay for Transposase Accessible Chromatin with high-throughput sequencing (ATAC-Seq). This procedure was performed as described in (Schep *et al.*, 2015) and (Buenrostro *et al.*, 2015) with slight modifications.

For the cell nuclei preparation, approximately 5 million cells (counted with an hemocytometer) were harvested (centrifugation at 500 g for 5 min, 4 *ºC*) and washed twice (centrifugations at 500 g for 5 min, 4 *ºC*) with 50 µl of cold sorbitol buffer (1.4 M sorbitol, 40 mM HEPES-KOH pH 7.5, 0.5 mM $MgCl_2$). We used Zymolyase 100-T (ZymoResearch) to remove the cell wall, and three different concentrations were tested before proceeding with the final experiments: 1 µl, 3 µl and 5 µl of zymolyase 5 U/µl. We incubated the cells with the corresponding amount of zymolyase in 50 µl of sorbitol buffer at 30 *ºC* for 30 min shaking at 300 rpm. Then cells were pelleted (500 g for 10 min at 4 *ºC*) and washed twice with 50 µl of cold sorbitol buffer (centrifugations at 500 g for 10 min, 4 *ºC*). Fresh pellets of fungal spheroplasts were brought to the Genomics Unit at the Centre for Genomic Regulation (CRG) for further transposase reactions and library preparations for ATAC-Seq. Briefly, nuclei were resuspended in 50 µL 1× TD buffer containing 2.5 µL transposase (Nextera, Illumina). The transposase reaction was conducted for 30 min at 37 °*C*. Library amplification and barcoding were performed with NEBNext Q5 Hot Start HiFi PCR Master Mix (New England Biolabs) using index primers, designed according to (Buenrostro *et al.*, 2015) at a final concentration of 1.25 µM. PCR was conducted for 12–13 cycles. Library purification was performed with AgenCourt AMPure XP beads (Beckman Coulter) and library size distribution was assessed using the Fragment Analyzer (AATI, Agilent) or the Bioanalyzer High Sensitivity DNA Kit (Agilent). The use of 3 µl of Zymolyase 5 U/µl was chosen as the optimal concentration for the experiments based on the visual inspection of the obtained profiles. ATAC-Seq libraries were quantified before pooling, and sequencing using the real-time library quantification kit (KAPA Biosystems). Paired-end sequencing was performed on a HiSeq 2500 (Illumina) with 50 cycles for each read.

All experiments were performed in three biological replicates. All library preparation and sequencing steps were performed at the Genomics Unit of the Centre for Genomic Regulation (CRG), Barcelona, Spain.

Sequencing reads quality control and visualization

We used FastQC v0.11.6 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) and Multiqc v. 1.0 (Ewels *et al.*, 2016) to perform quality control of raw sequencing data. Adapter trimming was performed by Trimmomatic v. 0.36 (Bolger, Lohse and Usadel, 2014) with TruSeq3 and Nextera adapters (for RNA-Seq and ATAC-Seq, respectively) using 2:30:10 parameters and the minimum read length of 30 bp. To visualize genomic/transcriptomic alignments and coverages we used the Integrative Genomic Viewer v. 2.3.97 (IGV)

(Robinson *et al.*, 2011).

RNA-Seq analysis

RNA-Seq read mapping and summarization was performed using the splice-junction aware mapper STAR v. 2.5.2b (Dobin *et al.*, 2013) with default parameters. For parental species, we mapped RNA-Seq data to the corresponding reference genomes, while for the hybrid strain we mapped raw data to the combined *S. cerevisiae x S. uvarum* reference genomes. Further, to assess the rates of reads originated from one species while mapped to another (*i.e.* cross-mapping, which possibly can bias the inference of gene expression levels) we employed two approaches: i) mapping the reads of each parental to the concatenated reference genome and then calculating the proportion of wrongly mapped reads to a different parental genome; ii) using the tool Crossmapper v. 1.1.0 (Hovhannisyan, Hafez, *et al.*, 2020), which simulates the data from both parental species, maps the reads to the concatenated genome and calculates the cross-mapping statistics. The reference genomes and genome annotations were obtained from Ensembl (release 93, (Zerbino *et al.*, 2018)) and www.saccharomycessensustricto.org (Scannell *et al.*, 2011) for SC and SU, respectively. For SU we merged ultrascaffolds and unplaced regions in one reference and converted GFF to GTF format using gffread v. 0.9.8 (Trapnell *et al.*, 2012) utility.

One-to-one orthologs between SC and SU were retrieved from *www.saccharomycessensustricto.org* (Scannell *et al.*, 2011). Differential gene expression and allele-specific expression were assessed using DESeq2 v. 1.18.0 (Love, Huber and Anders, 2014). For between-species comparisons we included the matrix of gene lengths to DESeq2 object to account for their differences. Additionally, for allele-specific expression analysis (within-hybrid comparison) we supplied the DESeq2 object with the matrix of gene lengths using *normalizationFactors(dds) <- lengths / exp(rowMeans(log(lengths)))*, allowing DESeq2 to account only for differences in gene lengths when calculating sizeFactors and ignoring the library size, since the read counts for alleles come from the same library. For a gene to be considered differentially/allele-specifically expressed, we used a threshold of |log2 fold change| (L2FC) > 1.5 and padj (adjusted p-value) < 0.01, unless specified otherwise.

Differentially expressed (DE) genes were used in Gene Ontology (GO) enrichment analysis as implemented in *Saccharomyces* Genome Database (Cherry *et al.*, 2012) to find functional enrichments in Biological Process, Molecular Function and Cellular Component GO categories. GO enrichment analysis for SU was done based on SC orthologous genes. To visualize the gene expression data we utilized ggplot2 v. 2_3.0.0 R library (Wickham, 2016).

ATAC-Seq analysis

Data generated by ATAC-Seq was mapped to the corresponding reference genomes using BWA v. 0.7.17-r1188 (Li and Durbin, 2009) with the MEM algorithm.

Initial mapping showed that ~15-18% of reads mapped to two regions of chromosome XII (450915-469179 and 489349-490611), which contain highly repetitive rRNA genes of SC. Thus, to remove the adverse effects in further analysis we have masked these two regions with bedtools maskfasta v. 2.27.1 (Quinlan, 2014).

PCR duplicates were marked using Picard MarkDuplicates v. 2.9.2 function (http://broadinstitute.github.io/picard). We used MACS2 v. 2.1.1 (Zhang *et al.*, 2008; Li and Durbin, 2009) to perform peak calling and the bedtools genomecov to generate bedgraph files of genome coverage by ATAC-Seq reads.

Bioconductor package DiffBind v. 2.4.8 (Ross-Innes *et al.*, 2012) was used to perform general quality control, and occupancy and affinity analysis of ATAC-Seq peaks. By occupancy analysis DiffBind finds overall peakset between replicates of a given biological condition and/or identifies consensus peaks between different biological conditions (i.e. parental peakset and peakset of the hybrid), while in affinity analysis it performs differential accessibility analysis of corresponding peaks, which is based on the DESeq2 workflow.

For comparing the peaksets between parentals and the corresponding homeologous chromosomes in the hybrid, we splitted the bam files of the hybrid into separate files for SC and SU chromosomes, using samtools v. 1.3.1 (Li *et al.*, 2009).

To perform differential accessibility (affinity) analysis within the hybrid, we first defined the orthologous/homeologous promoter regions as upstream, non-coding, genomic regions up to 1kb of length at each one-to-one orthologous locus. Defined promoter regions were usually shorter than 1kb since neighboring genes or chromosome borders were often encountered within that distance. We obtained bed files of promoter regions for each species using custom python scripts. Based on the bed files, we quantified the overlapping ATAC-Seq reads within promoter regions using bedtools multicov function. Further, differential accessibility analysis was performed using DESeq2 controlling for the length of regions.

We used bedtools closest, and custom python scripts to define, for each peak, the closest upstream and downstream genes within 1kb distance and for which the ATAC-Seq peak falls within the promoter region (suppl. Fig. 1).

Transcription factor footprinting

Besides defining open chromatin regions, we used ATAC-Seq data to perform transcription factor (TF) footprinting, to identify potential differences in TF binding sites occupancy between the parental and the hybrid. Position weight matrices for available *S. cerevisiae* TFs (n=176) were retrieved from the Jaspar database (Khan *et al.*, 2018). Footprinting was performed using HINT software of Regulatory Genomics Toolbox v0.11.4 (RGT) package (Gusmao *et al.*, 2014; Khan *et al.*, 2018; Z. Li *et al.*, 2019). Fungal organisms were added to HINT following the recommendations of the package developers. The *Motifanalysis* function of RGT package was used to match the motifs of fungal TFs with ATAC-Seq footprints. We used *differential* function of HINT to carry out differential TF binding site occupancy analysis and generate footprinting plots. The potential targets of differentially active TFs were identified using Yeastract platform (Teixeira *et al.*, 2018) by setting the Regulation filters to “DNA binding and expression evidence” to account only for targets genes with strong experimental evidence.

All custom scripts used in this study are available at https://github.com/Gabaldonlab/Hybrid_project.

Raw sequencing data of RNA-Seq and ATAC-Seq experiments were deposited in the Sequence Read Archive under the accession numbers SRR10246851-SRR10246868 and SRR10261591-SRR10261596, respectively.

8.4 Results

Limited transcriptional impact of hybridization

To assess the impact of hybridization on gene expression we used an RNA sequencing approach to profile the transcriptomes of diploid strains of *S. cerevisiae, S. uvarum,* and a *de novo* reconstructed diploid hybrid strain between these two species (see Materials and Methods). We repeated the experiment at 30 °*C*, a temperature within the optimal growth range of both species, and at 12 °*C*, which represents a cold shock, particularly for the non cryotolerant *S. cerevisiae* (Salvadó et al., 2011). This experimental design allowed us to directly compare transcriptional differences across genetic backgrounds (homozygous parentals and the hybrid), species (orthologous genes), homeologous chromosomes, and temperatures (see Fig. 8.1A) and therefore assess the relative impact of these factors on gene expression levels.

As recommended for robust inference of transcriptional levels (Liu, Zhou and White, 2014), we performed all experiments in three biological replicates, and sequenced over 30 million reads per replicate (see Materials and Methods).

The negligible level of cross-mapping between reads of the two species, as assessed by Crossmapper (Hovhannisyan, Hafez, *et al.*, 2020), suppl. file 1), and the independent mapping of parental RNA-Seq reads to both reference genomes (suppl. table 1B), allowed us to accurately assign reads to each parental sub-genome in the hybrid, and thus infer the relative expression of each of the two homeologous alleles. To additionally test whether these negligible cross-mapping rates can influence downstream results, we compared the read counts obtained from mapping parental data to the combined reference and the counts obtained by mapping parental data to corresponding parental genomes.

In case of both species, we observed Spearman’s correlations > 0.99, and that differential expression analysis with relaxed filters (|L2FC|>1, padj>0.05) did not show any gene affected by cross-mapping, verifying the accuracy of read assignments to corresponding species. Mapping statistics are shown in suppl. table 1, and quality control and reproducibility metrics for all samples are available in suppl. Fig. 2-4. Overall, we observed lower mapping rates of SU as compared to SC, which likely reflects the lower quality of the reference assembly for the former species.

For each of the 11 pairwise comparisons depicted in Fig. 8.1A we performed differential expression analyses (suppl. tables 2-13), tested for enrichment of functional GO terms among DE genes (suppl. tables 2-13), and assessed the correlation between the levels of expression (Fig. 8.1B).

Up-regulation of trehalose metabolism in *S. uvarum* sub-genome was also observed in the hybrid when it was exposed to 12 º*C* (suppl. table 12), which might be associated with adaptation of the hybrid to low temperatures (Pérez-Torrado *et al.*, 2015).

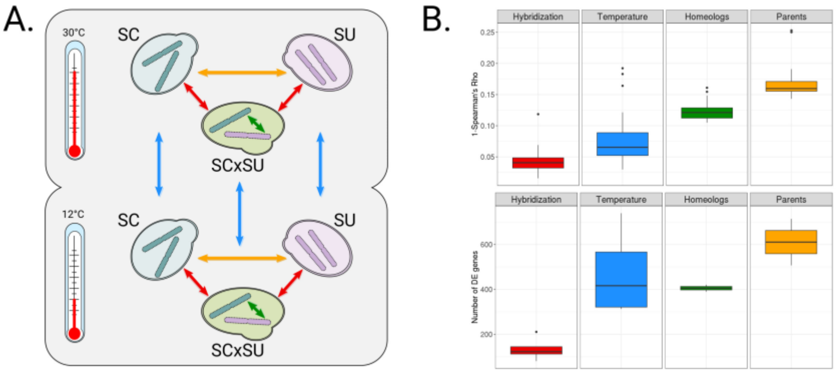


Fig. 8.1. (A) Experimental design of the study (see Materials and Methods). Arrows indicate comparisons of expression levels enabled by this design: across parentals (yellow), genetic backgrounds (red), homeologous genes (green), and temperature conditions (blue) **(B)** Overall transcriptomic changes assessed as 1-Spearman's Rho correlation (top row) and the number of DE genes (bottom row). "Hybridization" - comparisons between parentals and hybrid at both temperatures; "Temperature" - comparisons of all species at two different temperatures; "Homeologs" - comparisons between homeologous genes at both temperatures, "Parents" - comparisons between parentals at both temperatures. Colors correspond to the comparisons depicted in **(A)**. A more detailed comparison for each of the above categories is depicted in suppl. Fig. 6.

Most importantly, a comparison of the relative level of transcriptional differences across genetic backgrounds, temperatures, homeologous genomes, and species (Fig. 8.1B) shows that hybridization has a rather reduced transcriptional impact, being significantly lower than that observed for the temperature shift.

Overall, only 81 and 123 genes are differentially expressed when comparing hybrid and parental genetic backgrounds for *S. cerevisiae* and *S. uvarum*, respectively (suppl. tables 2 and 3) at 30 °C, which represents between 1% and 2% of the total gene repertoire of each species. In comparison, a temperature shift significantly alters the expression of 509 (7.1%) and 739 (11.5%) genes in these species, respectively. Additionally, we observed that differences of expression between orthologous genes in the two species were significantly larger than those observed between homeologous genes in the hybrid. This indicates that inter-species differences in terms of transcriptional landscape are attenuated rather than increased in the hybrid.

Low levels of allelic imbalance in yeast hybrid

We next explored allele-specific expression in the hybrid. Consistent with the largely conserved transcriptional landscape after hybridization, we found relatively low levels of allelic imbalance (i.e. significantly different expression levels of the two homeologous genes) (Fig. 8.1B). Specifically at 30 °*C*, 390 (~7.4% of the homeologous pairs, suppl. table 5) homeologous genes in the hybrid show allelic imbalance, from which 180 genes show higher expression of the SU allele, while 210 show higher expression of the SC allele. Thus, there is no strong preference for the hybrid to express one of the two sub-genomes. To identify whether the genes with allelic imbalance in the hybrid were a consequence of hybridization or were already differentially expressed when comparing the parental species (allele-specific expression inheritance), we compared the list of DE genes between parental species (suppl. table 4) with the list of imbalanced homeologous genes (Fig. 8.2).

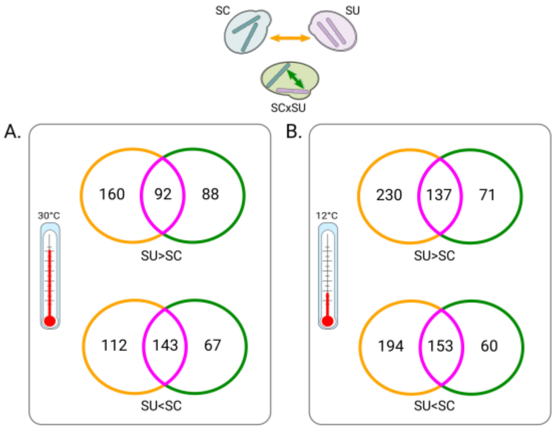


Fig. 8.2. Venn diagrams of between-parent (in yellow) versus within-hybrid (in green) comparison (depicted on the top) at **(A)** 30 °*C* and **(B)** at 12 °*C* . Intersections (in violet) indicate DE genes that appear in both conditions. Numbers indicate DE genes; colors of Venn diagrams correspond to the types of comparisons, as indicated with the arrows in the top scheme and consistent with Figure 8.1A (except for intersections); “>” and “<” symbols denote which homeologs or orthologs of a given species are expressed at significantly higher and lower levels, respectively (see suppl. tables 4, 5, 8 and 9 for lists of DE genes).

At 30 °*C* (Fig. 8.2A), 68% (143/210) and 51% (92/180) of the genes preferentially expressing the SC or SU alleles, respectively, were also found to have differential expression (with the same direction) in comparisons across species. Hence, this result indicates that the majority of genes with allele-specifically expressed genes in the hybrid are driven by inheritance of expression levels from parental species rather than resulting from the hybridization. Additionally, we found fewer genes that acquired allele-specific expression in the hybrid without being differentially expressed across species (88 and 67), as compared to genes that show no allele-specific expression despite being differentially expressed across the two species (160 and 112).

Overall similar trends were found at 12 *°C* (Fig. 8.2B). Collectively, these observations suggest that hybridization tends to attenuate, rather than exacerbate, differences in expression levels of parental orthologous genes. Finally, we also found that there is a small set (n=40) of temperature-dependent allele-specifically expressed genes (suppl. table 13), which is congruent with an early study (Li and Fay, 2017).

Overall conservation of genome-wide chromatin accessibility patterns after hybridization

We further investigated gene regulation differences upon hybridization by performing genome-wide chromatin accessibility analysis based on ATAC-Seq at 30 *°C* of the hybrid and the *S. cerevisiae* parental (see Materials and Methods, suppl. table 14). We compared ATAC-Seq profiles by performing peak calling and comparing the overlap between called peaks (i.e. inferred open chromatin regions) in the parental and the SC sub-genome of the hybrid (suppl. Fig. 5). After removing one outlier (see Materials and Methods), we found that replicate experiments showed large overlap of the called peaks (83% for both parent and hybrid replicates). Then, we compared peak sets of the SC parental with the corresponding sub-genome in the hybrid. This analysis showed that the state of chromatin accessibility is largely similar between the SC parental genome and the corresponding sub-genome of the hybrid (Fig. 8.3A).

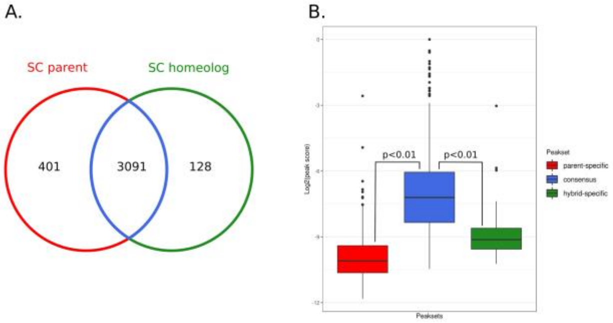


Fig. 8.3. Occupancy analysis. **(A)** Overlap between ATAC-Seq peak sets detected for SC genome and the corresponding *S. cerevisiae* sub-genome in the hybrid; **(B)** Comparison of distribution of DiffBind peak scores of parent-specific (red), consensus (blue) and hybrid-specific (green) peaks. p-values are calculated using Wilcoxon test.

From 3492 parental open chromatin regions, 3091 (88%, consensus peak set) are present in the hybrid and, conversely, 96% of the hybrid SC subgenome peaks are shared with the SC parental. Although they represent a small fraction, we did observed parent-specific (n=401) and hybrid-specific (n=128) accessible chromatin regions. However, we found that these specific peaks have significantly lower scores than the shared peaks (Fig. 8.3B), suggesting that some of these differences might represent false-positive peak calls.

Nevertheless, to assess whether parent- and hybrid-specific peaks in chromatin accessibility were driving observed transcriptional changes, we integrated ATAC-Seq and RNA-Seq data. For each parental- and hybrid-specific ATAC-Seq peak, we identified the nearest downstream gene for each strand, which potentially could be regulated by the open chromatin region identified by the peak. Only one gene near a parental-specific peak was found to be also over-expressed with respect to the hybrid context: the gene encoding pyridoxal-5'-phosphate synthase (YFL059W). Conversely, the gene coding for the NADH-dependent aldehyde reductase (YKL071W) was over-expressed in the hybrid and was sitting downstream of a hybrid-specific chromatin accessibility peak. The low fraction of peaks that are specific to each genetic background, their low scores, and the very low number of downstream genes that actually show differential expression, suggest that changes in chromatin accessibility upon hybridization have a very limited impact at the transcriptional level.

Further, we performed differential chromatin accessibility analysis (i.e. affinity analysis) within the consensus peak set, shared between the parental and the hybrid. Even when using liberal thresholds (L2FC > 1 and FDR<0.01), we found only two differentially accessible peaks in this consensus set: namely regions at the chromosome XII 682106-682709 (more open in the SC parent, L2FC=1.3) and IX 391660-392121 (more open in the hybrid, L2FC=1.32). We combined this result with RNA-Seq and visualized the integrated data (Fig. 8.4).

In the first region (Fig. 8.4A), the gene YLR271W that is downstream of the peak is not differentially expressed between the parent and hybrid. In contrast, the second peak (Fig. 8.4B) coincides with the significantly higher expressed gene FLO11 (YIR019C, L2FC = 3.44, padj<0.01) in the hybrid, as compared to the parent. However, in this case, the peak entirely overlaps the gene, and therefore it is unlikely that it regulates its expression. Thus, differential levels of chromatin accessibility do not seem to drive the few differences observed between hybrids and parental genetic backgrounds.

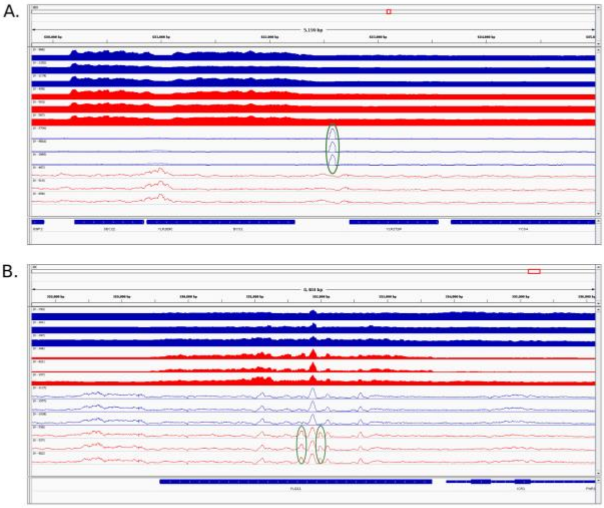


Fig. 8.4. IGV screenshots combining RNA-Seq and ATAC-Seq datasets for SC and its counterpart in the hybrid. **(A)** The first identified differentially accessible region; **(B)** The second identified differentially accessible region. Blue tracks correspond to SC in parental background, red tracks correspond to SC in hybrid background. Filled tracks correspond to RNA-Seq data, contour tracks correspond to ATAC-Seq data. Regions with higher coverage are highlighted in green circles. The last track represents genomic features. Further description is given in the main text.

Next, taking advantage of the high sequencing depth of our ATAC-Seq data, we also assessed the changes in TF activity (as defined by Li et al., (2019)) in a genome-wide manner. The results show (Fig. 8.5) that only 8 out of 176 *S. cerevisiae* TFs have significantly ($p<0.05$) changed their activity levels upon hybridization. In most cases, the differences in activity are moderate, with the notable exception of ARG81, which mediates the arginine-dependent repression of arginine biosynthesis genes and the activation of arginine catabolic genes.

We further identified the potential target genes for each of the deregulated TFs (suppl. table 15), and compared the target genes with DE genes upon hybridization. From the 189 target genes of these 8 TFs, only 2 genes corresponded to genes differentially expressed in the same direction as their TF - YHL040C (gene encoding for iron-siderochrome transporter) and YLR346C (encodes for a protein of unknown function), were up-regulated in the hybrid genetic background.

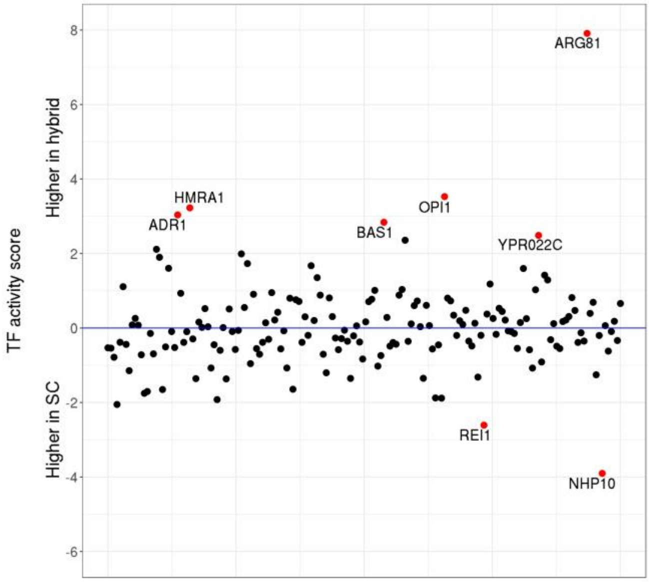


Fig. 8.5. TF activity scores. Relative activity levels between SC parental and the hybrid counterparts. Red dots highlight the TFs which significantly (p<0.05) changed their activity levels upon hybridization.

Both of these genes are regulated by BAS1, a TF involved in regulating expression of genes of the purine and histidine biosynthesis pathways. Overall, our results suggest that upon hybridization the changes in TF activities are subtle and largely do not correlate with the patterns of differential gene expression observed by RNA-Seq.

Chromatin accessibility patterns within the hybrid are weakly correlated with allele-specific gene expression

Finally, we compared the chromatin accessibility profiles of SC and SU homeologous regions within the hybrid. First, we defined the homeologous regions between two sub-genomes as maximum of 1 kb upstream regions of each 1-to-1 orthologous gene. Further, we quantified the ATAC-Seq coverage for these regions, and performed differential accessibility analysis using DESeq2, controlling for the length differences of the regions. This analysis identified 59 and 75 genomic regions that were significantly more open or less open, respectively in the SU sub-genome as compared to the SC sub-genome (|L2FC| > 1, padj<0.01). By comparing these data with the results of allele-specific expression, we found that 8 out of 180 preferentially expressed SU homeologs coincided with significantly more open SU regions, and 9/210 preferentially lower expressed SU homeologs corresponded to significantly less open SU regions. These results are in agreement with a previously published study, which identified a low correlation between changes in nucleosome positioning and gene expression levels in yeasts (Tirosh, Sigal and Barkai, 2010).

8.5 Discussion

Fungi, and in particular Saccharomycotina yeasts have been shown to be prone to hybridization, with an increasing number of hybrid species that are highly successful in certain niches and have industrial or clinical relevance (Pryszcz *et al.*, 2014; Krogerus *et al.*, 2017; Mixão *et al.*, 2019). Hybridization has also been shown to be at the root of entire clades - e.g. the post-whole genome duplication clade comprising *Saccharomyces* and related genera has been shown to result from a hybridization event (Marcet-Houben and Gabaldón, 2015). Thus, rather than representing evolutionary dead-ends, fungal hybrids might be highly successful and long-lived. This implies that fungal species which form a hybrid organism must overcome molecular differences and potential incompatibilities which evolved through the evolutionary history of the parentals. On the relatively well-studied genomic level, fungal hybrids, and in particular those from the *Saccharomyces* genus, tend to undergo genomic rearrangements, including loss of heterozygosity, gene conversion, partial or full chromosome loss, among others, that help to overcome incompatibilities and stabilize genomes, which results in genome mosaicism.

In our study, we assessed how the two distantly related *Saccharomyces* species cope with hybridization at the levels of the transcriptome and chromatin landscapes.

Collectively, our results show that, despite genome merging of extremely diverged species, hybridization has a comparatively smaller effect on the transcriptome than a shift from temperate to cold temperature. Moreover, we found that most loci express the two homeologous alleles in similar proportions, and those genes that show allele-specific expression largely overlap with those whose orthologs in the two parental species already show different levels of expression, suggesting that most of the allele-specific expression derives from already existing inter-species differences. Furthermore, homeologous genes within the hybrid showed less differences than orthologous genes in the two parental species, indicating that, rather than being exacerbated inter-species differences in expression are attenuated upon hybridization. This is consistent with an earlier study on a natural hybrid of the genus *Epichloë* (Cox *et al.*, 2014). However, as our model involves a newly formed hybrid, our study clarifies that the attenuation of the differences is an immediate effect of hybridization and not the result of adaptation through evolution of the hybrid lineage.

Additionally, Cox et al., proposed that with an increase in genome divergence between the parentals, which ~5% in *Epichloë* (Campbell *et al.*, 2017), the magnitude of transcriptome shock will increase accordingly, while our results demonstrate that despite a large evolutionary distance of 20% of nucleotide divergence in coding regions (Kellis *et al.*, 2003) between SC and SU the consequences of hybridization are still buffered.

Moreover, in contrast to the *Epichlöe* study that compares haploid parentals with a diploid hybrid, our study compares diploid parental strains and diploid hybrids and thus avoids any potential misleading effect resulting from a ploidy change.

The absence of a large impact of hybridization in gene expression in fungal hybrids is in stark contrast with what has been reported in animals or plant studies (McManus *et al.*, 2010; Yoo, Szadkowski and Wendel, 2013; A. Li *et al.*, 2014; Wu *et al.*, 2018; Zhang *et al.*, 2018), where widespread changes in expression following hybridization have been observed. For instance, in newly resynthesized allotetraploid *Brassica napus* (Wu *et al.*, 2018), 30.4% of the genes showed expression changes upon hybridization compared to its diploid parentals *B. rapa* and *B. oleracea*, and over 90% of the deregulated genes were down-regulated in the allotetraploid compared to the parents. Additionally, 36.5% of homeologous pairs within the hybrid *B. napus* displayed differential expression towards either of the alleles, with a slight preference to the *B. rapa* parental. Similarly, in allopolyploid cotton *G. arboreum (A2) x G. raimondii (D5)*, 22-30% of parental genes showed differential expression in the hybrid, as compared to the parentals (Yoo, Szadkowski and Wendel, 2013). The study of allelic imbalance of synthetic hexaploid wheat showed that 24.1% of the identified homeologous genes were imbalanced in the hybrid, and this difference in expression could not be attributed to pre-existing expression divergence between the parentals (A. Li *et al.*, 2014). Finally, a recent study on the diatom microalgae *Fistulifera solaris,* showed that ~61% of homeologous genes displayed allelic imbalance (Nomaguchi *et al.*, 2018).

For *Drosophila* hybrids different amplitudes of transcriptome misexpression have been previously reported, which depend on the level of genetic divergence of the parental species. For instance, a recent transcriptomic study of *D. mojavensis* and *D. arizonae* (diverged 0.6-1 million years ago) and their hybrid showed that 12% of genes in the hybrid are differentially expressed, as compared to the parents (Lopez-Maestre *et al.*, 2017). This is much larger than the fraction of DE genes in this study, despite the much lower genetic distance in the *Drosophila* species. The same study showed 8% of genes between parentals have diverged expression. That is, in that case differential expression between homeologous genes in the hybrid was more abundant than between orthologous genes in the two parental species, which is the contrary of what we have observed for *Saccharomyces* hybrid in this study. On the other hand, the comparison between more diverged fly species, namely *D. melanogaster* and *D. sechellia* (diverged ~1.2 million years ago), identified 78% of DE genes between parents (McManus *et al.*, 2010). Interestingly, a recent study of hybrid chicken breeds (intra-species hybrids) showed a tissue-dependent rates of gene expression divergence: while it was ~15% of genes in liver, gene expression divergence between parental breeds in brain was as low as 0.8% of genes (Gu *et al.*, 2019).

It must be noted that comparing results across species and studies is difficult. These studies were performed using different technologies and data analysis methods, which makes direct comparisons problematic. For instance, Wu and colleagues (Wu *et al.*, 2018) used FPKM expression values and applied a filter of FDR≤0.05 and absolute log2 fold change ≥ 1 for a gene to be considered as differentially expressed; Yoo et al., (2013) used raw read counts for differential expression analysis and a filter of adjusted p-value < 0.05 with no filtering on fold-change; (Gu *et al.*, 2019), applied an absolute fold-change≥1.25 and FDR<0.5 for assigning DE genes in chicken hybrids breeds; and Lopez-Maestre et al. (2017) used FDR<0.01 and log2 fold change > 1.5 for assigning DE genes in *Drosophila*., that are the filters also used in our study. In order to assess how data filtering can influence our results, we applied a set of more liberal filters for finding DE genes upon parental-hybrid transition. With padj (FDR)<0.05, |log2 fold change|>1, and mean normalized expression levels > 10 (which in fact takes into account only expressed part of the transcriptome, limiting transcriptome-wide inferences) we obtained on average ~4.6% and ~9% of DE genes for SC and SU, respectively, across temperatures. This shows that even with relaxed filters, transcriptome shock in our yeast hybrid is lower than in plants and animals. Thus, methodological differences notwithstanding, the different animal and plant studies seem to agree in reporting a large transcriptomic impact of hybridization, as well as large levels of allele-specific imbalance, whereas the fungal studies consistently report more moderate effects.

We further assessed the impact of hybridization on another level: that of chromatin accessibility. Here, consistent with the low level of differences in gene expression, we found minor differences in terms of chromatin accessibility and TF activity between the hybrid and the *S. cerevisiae* parental. Admittedly, subtle differences in TF expression might have significant biological effects. However, our results suggest that the few observed differences in chromatin accessibility and TF activity are not driving the few observed differences in gene expression levels.

Based on our results and those from other previous studies, we hypothesize that unicellular fungi and multicellular plants and metazoans respond fundamentally different to hybridization in terms of transcriptional response which may explain why hybridization is so common in fungi, and, as compared to plants and animals, it can encompass larger genetic distances (Morales and Dujon, 2012). What molecular phenomena govern these different responses to hybridization? One could argue that differences in magnitudes of transcriptomic shock in fungi and metazoans and plants can be attributed to differences in mechanisms regulating gene expression. Though the general and fundamental principles of transcriptional regulation are largely conserved across eukaryotes, the complexity of gene regulation in plants and animals is more sophisticated than in fungi (Rando and Chang, 2009; Hahn and Young, 2011; Lelli, Slattery and Mann, 2012). For example, plants and animals possess a richer repertoire of chromatin modification regulators, as compared to yeasts, which provide them with additional layers of regulation and a more sophisticated fine tuning of expression levels (Rando and Chang, 2009).

Additionally, fungi and yeasts in particular are prone to genomic rearrangements, ranging from small indels to large scale copy number variations, inversions, translocations and duplications (Albertin and Marullo, 2012; Plissonneau, Stürchler and Croll, 2016; Möller and Stukenbrock, 2017; Möller *et al.*, 2018; Steenwyk and Rokas, 2018). Not only these genomic alterations are compatible with fungal viability, but also, inversely, can promote fitness and adaptability to different niches (Selmecki, Bergmann and Berman, 2005; Selmecki *et al.*, 2009; Croll, Zala and McDonald, 2013; Jonge *et al.*, 2013; Gabaldón and Carreté, 2016; Mixão and Gabaldón, 2018). Hence, one could expect that, following hybridization, two complex gene regulatory systems, such as that of animals and plants, are more likely to experience larger levels of incompatibilities and perturbations, as compared to simpler and more versatile regulatory systems, such as those of yeasts.

In this context, Cox et al., (Cox *et al.*, 2014) introduced the concept of "modulon" which encompasses all gene regulatory mechanisms, including *cis*- and *trans*- regulation, post-transcriptional regulation, TFs, epigenetics, etc. Differences in the levels of expression of orthologous genes in different species arise due to differences in the species' modulons, which have evolved independently for some time. Upon hybridization, several regulatory scenarios can take place: (i) modulons of the two species have no or little crosstalk with each other because they are too divergent; (ii) modulons are largely similar and compatible with each other, resulting in a so- called homeolog expression blending; or (iii) modulons of the two species preferentially target one of the alleles. Importantly, these regulatory outcomes can coexist, affecting different portions of the transcriptome, which can be quantitatively assessed. In the first scenario, the genes from the parental species would inherit their expression levels in the hybrid with no subsequent expression alterations. This outcome accounts for the majority of genes in our study - ~92% and ~89.3% of orthologous genes at 30 °*C* and 12 °*C*, respectively (non-DE orthologs + violet parts of Fig. 8.2). The second scenario will result in diminished differences in homeologous expression levels as compared to differences across species. In our study, this could account for ~5.24% and ~8.2% of orthologous genes at 30 °*C* and 12 °*C*, respectively (yellow parts of Fig. 8.2). In the third case, homeologous genes will acquire divergence in gene expression that was not observed in parentals, which represents the transcriptomic shock caused by hybridization. In our study, this accounts for ~2.9% and ~2.52% of homeologous genes at 30 °*C* and 12 °*C*, respectively (green parts of Fig. 8.2).

Altogether, our study suggests a conservative and restricted impact of hybridization at the transcriptomic and chromatin profiles in hybrid yeast, which can be largely attributed to the absence of regulatory crosstalk between highly diverged fungal modulons. We hypothesize that the moderate impact that hybridization has on the levels of chromatin accessibility and gene expression is at the root of the strong ability for successful hybridization in yeasts and other fungi. Further research involving diverse taxonomic groups of fungi is required to address this hypothesis in order to disentangle the role transcriptome and chromatin profile buffering in fungal hybridization.

Acknowledgments

We thank Dr. Susana Iraola (Comparative Genomics group, BSC/IRB, Barcelona, Spain) for her support in the laboratory experiments. We also thank the team of rgt-hint software developers for technical support and valuable recommendations in ATAC-Seq footprinting analysis.

Data Availability Statement

Raw sequencing data of RNA-Seq and ATAC-Seq experiments were deposited in the Sequence Read Archive under the accession numbers SRR10246851-SRR10246868 and SRR10261591-SRR10261596, respectively.

Supplementary information can be found at https://www.frontiersin.org/articles/10.3389/fgene.2020.00404/full#supplementary-material

Funding

This work was funded in part by the European Union’s Horizon 2020 research and innovation programme 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 Programme / Generalitat de Catalunya; from the Catalan Research Agency (AGAUR) SGR857, and grants from the European Union’s Horizon 2020 research and innovation programme under the grant agreement ERC-2016-724173. TG also receives support from a INB Grant (PT17/0009/0023 - ISCIII-SGEFI/ERDF).

9 CROSSMAPPER: estimating cross-mapping rates and optimizing experimental design in multi-species sequencing studies

Hovhannisyan, Hrant*, Ahmed Hafez*, Carlos Llorens, and Toni Gabaldón. "CROSSMAPPER: estimating cross-mapping rates and optimizing experimental design in multi-species sequencing studies." *Bioinformatics* 36, no. 3 (2020): 925-927. DOI: 10.1093/bioinformatics/btz626

***** - equal contribution.
H. Hovhannisyan has designed the project, written the initial data analysis pipeline, documented, tested and maintains the software. A. Hafez has implemented the final pipeline and its automation

9.1 Abstract

Motivation

Numerous sequencing studies, including transcriptomics of host-pathogen systems, sequencing of hybrid genomes, xenografts, mixed species systems, metagenomics, and metatranscriptomics, involve samples containing genetic material from divergent organisms. A crucial step in these studies is identifying from which organism each sequencing read originated, and the experimental design should be directed to minimize biases caused by cross-mapping of reads to incorrect source genomes. Additionally, pooling of sufficiently different genetic material into a single sequencing library could significantly reduce experimental costs but requires careful planning and assessment of the impact of cross-mapping. Having these applications in mind we designed Crossmapper, the first to our knowledge tool able to assess cross-mapping prior to sequencing, therefore allowing optimization of experimental design.

Results

Using any combination of reference genomes, Crossmapper performs read simulation and back-mapping of those reads to the pool of references, quantifies and reports the cross-mapping rates for each organism. Crossmapper performs these analyses with numerous user-specified parameters, including, among others, read length, read layout, coverage, mapping parameters, genomic or transcriptomic data. Additionally, it outputs the results in highly interactive and publication-ready reports. This allows the user to perform multiple comparisons at once and choose the experimental setup minimizing cross-mapping rates. Moreover, Crossmapper can be used for resource optimization in sequencing facilities by pooling different samples into one sequencing library.

Availability and implementation

Crossmapper is a command line tool implemented in Python 3.6 and available as a conda package, allowing effortless installation. The source code, detailed information and a step-by-step tutorial is available at our GitHub page https://github.com/Gabaldonlab/crossmapper.

Contact
tgabaldon@crg.es

9.2 Introduction

There are various biological problems addressed by next-generation sequencing (NGS) in which the samples contain genetic material from multiple species. These include, but are not limited to studies involving host-pathogen interaction (Westermann, Barquist and Vogel, 2017), symbiont-host or microbial interaction (González-Torres *et al.*, 2015; Burns *et al.*, 2017), metagenomics (Quince *et al.*, 2017), or hybrid organisms (Metzger, Wittkopp and Coolon, 2017). A challenging step in these experimental setups is to assign each sequencing read to the corresponding source organism, which is usually done by mapping the reads to the set of reference genomes (Wolf *et al.*, 2018). A similar strategy is applied in allele-specific expression (ASE) studies in the case of phased reference genomes (Yuan and Qin, 2012). Successful read separation depends on numerous factors, including mainly read length, read layout, similarity of sequenced genomes and different mapping parameters. Thus, if these parameters are not carefully planned, downstream analyses can be biased by cross-mapping of reads to non-corresponding references. For example, in a human-*Salmonella* interaction study it was observed that ~1.44% of total reads map equally well (multi-mapped) to both reference genomes (Westermann and Vogel, 2018). While the amount of erroneously mapped reads can be low for highly divergent species, in metagenomics (Petersen *et al.*, 2017) and ASE studies, erroneously mapped and multi-mapped reads constitute the majority of the data (Yuan and Qin, 2012). Despite the importance of sequencing design in aforementioned studies, today there are no computational tools to assist in their planning so that optimal results are obtained.

To overcome this, we developed Crossmapper – a pipeline assessing, prior to sequencing, the potential rates of multi-mapping and erroneous mapping for various combinations of sequencing parameters and any number of reference sequences.

9.3 Workflow and implementation

Crossmapper proceeds as follows (Fig. 9.1A). It first takes as input any number of reference genomes and allows to simulate DNA and RNA reads in a wide range of experimental setups. This step is performed by wgsim (Li *et al.*, 2009) with the possibility to define different parameters such as read length, error rates, outer distance, among others. Crossmapper allows to simulate many different sequencing configurations at once. The user can specify genome annotations to limit read simulations from specific parts of the genomic regions (i.e. for transcriptomic or exome sequencing studies).

After read simulation, Crossmapper concatenates fastq files from different organisms and maps the reads back to a concatenated set of reference genomes. By default Crossmapper uses BWA-MEM (Li and Durbin, 2009), and STAR (Dobin *et al.*, 2013) for mapping DNA and RNA data, respectively. However, we also implemented the --*mapper-template* option allowing to use any desired mapping software with custom parameters by supplying the configuration file to the Crossmapper (a documentation for creating a configuration file is given in the GitHub page). The final bam file for each read length and layout contains alignments of all simulated reads collectively mapped to all source reference genomes. Since simulated data preserve information regarding the source genome and exact location, Crossmapper can calculate the rate of multi-mapped and erroneously mapped reads for all source genomes.

After the quantification step Crossmapper produces an extensive html report, which includes several interactive, publication-ready plots summarizing mapping rates, as well as tables with detailed mapping statistics for each experimental configuration. Based on this report users can decide the optimal experimental and mapping parameters prior to the actual sequencing. In addition, coordinates of cross-mapped reads are reported so these regions can be filtered, if necessary, in downstream analyses.

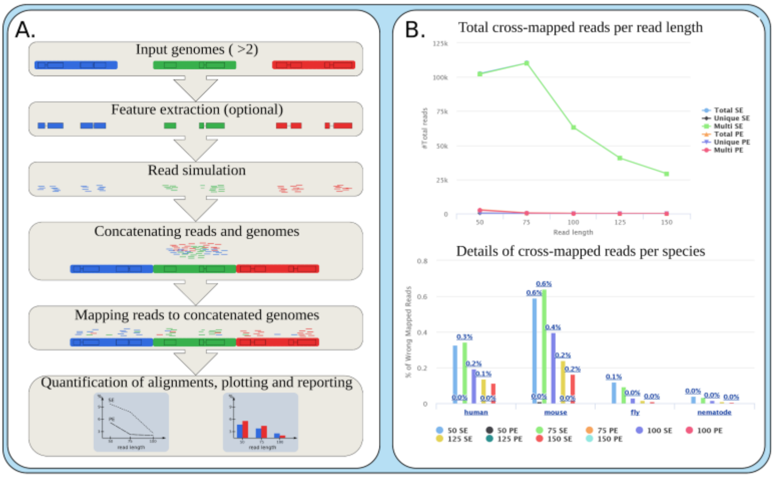


Fig. 9.1. (**A)** The general workflow of Crossmapper (see main text for details). (**B)** An example of Crossmapper output.

9.4 Usage case

Several examples of Crossmapper usage are available in the GitHub site of this tool. Here, we explain how to use of Crossmapper to optimize resources by pooling of genetic material of different organisms into a single sequencing library. Indeed, the cost of sequencing has dropped dramatically in the past decade (Goodwin, McPherson and McCombie, 2016), largely due to throughput increase. However, the costs for library preparation do not follow the same trend and often constitute a financial bottleneck. A simple pooling of genetic materials of different species into one library could save a substantial amount of resources, provided reads from different sources could effectively be separated computationally. This has to be carefully planned to avoid aforementioned biases in downstream analyses. Crossmapper can achieve this task in a single run. Below is an example of sequencing design optimization for pooling genetic material of widely analyzed organisms – human, mouse, fly and nematode – in a single library.

Command syntax

```
crossmapper DNA -t 8 -gb -rlay both -g homo.fasta mus.fasta dros.fasta caeno.fasta -gn human mouse fly nematode -N 2500000 2500000 2500000 2500000 -rlen 50,75,100,125,150 -r 0.01
```

This command lets Crossmapper to simulate 2.5 million DNA reads per organism at 50, 75, 100, 125 and 150 read lengths at both single- and paired-end layouts, map the data to the pool of reference genomes (obtained from

Ensembl (Zerbino *et al.*, 2018)) and report mapping rates for all sequencing configurations (Fig. 9.1B). Using Intel Xeon 3.5GHz, 64GB of RAM and 8 cores the analysis takes ~11 hours. In this case of very large reference genome (ca 6.5 GB) and 10 mapping jobs, the main bottleneck for the speed of the analysis is genome indexing and read mapping, which collectively takes ~8 hours.

The results of this analysis (suppl. file 1) demonstrate that by pooling the DNA of the 4 species reads can be effectively separated by mapping. However, single-end sequencing produces relatively high rates of multi-mapping (maximum 4.95% and 6.06% for 150 and 50 bp, respectively) and erroneous mapping (maximum 0.16% and 0.52%, for 150 and 50 bp, respectively) which potentially can bias differential expression or variant calling analysis. On the other hand, paired-end sequencing with 75 bp reads significantly reduces mutli- and erroneous mapping (0.01% and 0%, respectively) rates. Thus, the pooling strategy with 2x75 bp reads can be the most efficient balance between accuracy and sequencing cost. Repeating this test with a higher number of reads (40, 30, 20 and 10 mln reads for human, mouse, fly and nematode, respectively), showed similar rates of cross-mapping (suppl. file 2), which indicates that low-coverage simulations are sufficient to properly estimate cross-mapping rates.

9.5 Conclusion

Crossmapper allows to design numerous types of NGS experiments that share a common feature of sequencing several organisms as one sample. Crossmapper is easy to install and use. It is highly customizable and outputs the results in intuitive, interactive and publication-ready reports. We believe that Crossmapper will benefit both research and industrial communities by helping to optimize sequencing strategies and available resources.

Supplementary information can be found at https://academic.oup.com/bioinformatics/article/36/3/925/5544929#supplementary-data

Acknowledgements

TG group acknowledges support from the Spanish Ministry of Economy, Industry, and Competitiveness (MEIC) for the EMBL partnership, and the grant ‘Centro de Excelencia Severo Ochoa’ SEV-2012-0208 cofounded by European Regional Development Fund (ERDF); from the CERCA Programme / Generalitat de Catalunya; from the Catalan Research Agency (AGAUR) SGR857, and grants from the European Union’s Horizon 2020 research and innovation programme under the grant agreement ERC-2016-724173, and the Marie Sklodowska-Curie grant agreement No H2020-MSCA-ITN-2014-642095. The group also receives support from a INB Grant (PT17/0009/0023 - ISCIII-SGEFI/ERDF).

10 Summarizing discussion

In this PhD we addressed several major aspects of two interrelated phenomena - interactions of *Candida* pathogens with the human host, and the emergence of virulence in these yeasts. Additionally, we developed novel experimental and computational approaches to facilitate the investigation of these two processes, but which are of more general applicability. All these studies are described in different chapters of this thesis and all of them are built around the same methodological backbone: comparative transcriptomics by means of RNA-Seq. Every chapter has its own discussion section. Thus, here I will summarize and discuss broader implications of this PhD project and present my vision on future directions of the topics addressed in this thesis.

10.1 Novel insights into the human-*Candida* interaction mechanisms

Experimental design as a crucial prerequisite for transcriptomic studies

In Chapter 4 we investigated host-pathogen interactions between the four most widespread *Candida* species and human vaginal epithelial cells. We tackled this complex issue by using a recently developed methodology of dual RNA-Seq of the host and pathogen throughout the course of infection. Though it might be seen as a very technical aspect, the experimental design of this project, and in general any complex inferential transcriptome study, deserves special attention. This is particularly necessary considering that transcriptomics, unlike some other omics approaches, deals with a highly dynamic and unstable biological phenomenon as it is gene expression, and not only the expression of one gene but all of them simultaneously, at a specific time, in a specific condition. The list of “specifics” can be continued to include many other characteristics, which are often of technical nature – specific lab personnel, specific equipment, specific sequencing technology, etc. (SEQC/MAQC-III Consortium, 2014). All these parameters inevitably influence the obtained transcriptome profiles, usually in unpredictable ways. From a technical perspective, there are additional issues, such as replication, sequencing configuration, among others, that need to be considered in transcriptomics studies, especially inferential ones (Liu, Zhou and White, 2014; Conesa *et al.*, 2016; Yu, Fernandez and Brock, 2017). However, even if experiments are done with a minimal technical variability, the major aim of a transcriptomic project can be ruined if the experimental setup is flawed - if control experiments (or at least control genes) are overlooked, for instance, it is simply impossible to dissect the effect of a specific biological condition on a transcriptome profile if another independent condition acts at the same time. The combination of all above mentioned issues must be thoroughly assessed before proceeding with the actual experiments, because they are crucial for a meaningful and potentially successful transcriptome-based study (Hovhannisyan and Gabaldón, 2019).

The experiment in this project is complex, comprising an *in vitro* time-course model of vaginal candidiasis. Importantly, our model did not include the host immune system, mimicking a severely immunocompromised state. This design has two advantages: 1) it lets the fungi to fully deploy their pathogenic arsenal and 2) it allows to dissect the pure epithelial transcriptome response without being convoluted with transcriptome data from immune cells. During the course of infection, RNA from both counterparts was isolated at specific-time points - 3 hours post infection (hpi), 12 hpi and 24 hpi. We have chosen these time points based on the hallmark events of *Candida* infection - adhesion, invasion and damage - which correspond to 3, 12 and 24 hpi, respectively (Wächtler *et al.*, 2011; Wächtler, Wilson and Hube, 2011). However, it is important to mention that timing of these events is based on *C. albicans* and we simply do not know when these events happen for other species. In fact, this fundamental question of when these hallmarks of infection take place in different *Candida* pathogens and how they are biologically comparable, in my opinion, deserves a separate PhD project.

From other more technical perspectives, our experimental design followed the best practices (Conesa *et al.*, 2016) for transcriptomic studies - we used control samples allowing to dissect the effect of culturing media from the one of infection process; all samples had biological replicates and sufficient sequencing configuration and depth. Though it has to be mentioned that a considerable data analysis effort was done to integrate the RNA-Seq data of the follow-up experiments with our initial data due to observed batch effects. This, once again, highlights the importance of study design in complex studies, particularly when follow-up experiments are done to test proposed hypotheses.

Candida yeasts show highly specific transcriptome profiles

Equipped with large-scale and high quality RNA-Seq data, we were able to disentangle the major question which was initially posed - how do human-*Candida* interactions vary across fungal species. Importantly, our experimental approach allowed us to thoroughly study both sides of interacting counterparts.

First, we show that, on a broad scale, the transcriptional response through the whole course of infection of the four *Candida* species are remarkably different from each other. Each species differentially regulated a distinct set of genes when comparing the same time points. One the other hand, one could argue, as mentioned above, that the time points of different species might not be equivalent. However, this would only strengthen the idea that virulence programs differ between these species. From an evolutionary standpoint this implies that on the transcriptional level *Candida* species have developed unique adaptive mechanisms for living within and invading the human host. It is unlikely to be a result of divergent evolution, where their common pathogenic ancestors gradually have lost some pathogenic features. A more plausible scenario is that pathogenicity of these species in a broad sense (and from the host standpoint) is a result of convergent evolution, where phylogenetically diverse species have developed different mechanisms leading to a similar outcome for the host. In fact, phylogenomics suggests a similar scenario, showing that these pathogens are surrounded by non-virulent species, implying that pathogenicity towards humans (at least to epithelial cells) emerged independently in these yeasts (Gabaldón, Naranjo-Ortíz and Marcet-Houben, 2016).

Another important result in the context of fungal pathogenicity is that the global transcriptome profiles of virtually all four species were very similar between growth with human cells and growth on a rich culture medium, which also was observed previously for *C. albicans* (Liu *et al.*, 2015). Nevertheless, for all species, we were able to identify a relatively small number of genes which were exclusively differentially expressed due to the presence of the host. The functions of these genes are yet to be determined in future studies. These observations suggest that the virulence programs of these fungi mostly overlap with mechanisms governing their normal metabolism and functioning, underlying the importance of the concept that microbial pathogenesis *per se* must be interpreted through the lens of interaction of both counterparts - the host and the pathogen combined (Pirofski and Casadevall, 2008).

These two observations are not only important for understanding the biology of these pathogens, but also they can potentially drive some conceptual changes with respect to strategies for candidiasis treatment and *Candida* research in general. The fact that the transcriptional responses of these fungi are significantly different from each other means that many aspects of medical interventions need to be differentiated among these species. This also highlights the importance of precise diagnostics, since the infecting agents of candidiasis are often hard to diagnose accurately, especially when using conventional methods (Consortium OPATHY and

Gabaldón, 2019). Further, the differential approach might be necessary for the development of effective and specific antifungal drugs, especially considering the rise of non-albicans *Candida* species and emergence of drug resistance (Ksiezopolska and Gabaldón, 2018; Taei, Chadeganipour and Mohammadi, 2019).

The observation of similar transcriptional profiles of all studied *Candida* pathogens in the culture medium and in contact with the host, in my opinion, raises a thought that prioritization between basic and potentially applied research needs to be balanced. Many *Candida* studies are focused on identifying genes, pathways and mechanisms that play a direct role in pathogenesis, and thus can be used, for example, as a target for therapeutics. With our current understanding of similarity between normal fungal growth and fungal pathogenic programs, the research aimed for in depth understanding of basic biology of these pathogens in a long run might pay off as a more effective approach to combat these pathogens. However, this top-down transition would require efforts not only of the scientific community, but importantly also those of policy-making and science funding bodies.

Human mitochondria - Éminence grise of host defence

The analysis of human transcriptome profiles during infections revealed several important phenomena. First, epithelial cells have a conserved initial response to the four phylogenetically different fungal pathogens. Such kind of a uniform behavior of the human host is, in a way, expected. After all, the epithelium is the first barrier between human and the environment, which hosts different microbes, including fungi (Kim *et al.*, 2010; Dambuza and Brown, 2018; Richardson, Ho and Naglik, 2018). From an evolutionary perspective this would be the most efficient and parsimonious way of coping with a large number of various microbes. That being said, we currently have a very fragmented understanding of how this conserved response is deployed. In fact, only the terminal stages of this epithelial response are investigated to some extent, while the intermediate molecular processes are elusive. For example, the initial recognition of *Candida* depends on several receptors, such as TLR, C-type lectins, dectin-1 and EphA2 (Weindl *et al.*, 2007; Bahri *et al.*, 2010; Moyes *et al.*, 2010; Swidergall *et al.*, 2018). The final stage of downstream signaling of these receptors can lead to activation of three main pathways - MAPK, NF-κB and PI3K - involved in several basic processes such as cell proliferation and innate immune activation (Richardson, Ho and Naglik, 2018). However, these steps were mainly investigated for *C. albicans* and we still do not know which molecular events happen in between these initial and final stages of interaction.

In our work, we unraveled a novel molecular mechanism which underlies the uniform response to studied *Candida* species and to some extent closes the gap between the initial and terminal stage of epithelium-*Candida* interaction. Unexpectedly, the central players of this mechanism are human mitochondria, which recently have been recognized as central hubs of the innate immune system (Mills, Kelly and O'Neill, 2017). We have experimentally shown that upon challenging with all studied *Candida* species, already at the initial time point, human mitochondria undergo extensive activation of mtDNA-encoded genes, mtROS production, morphological changes, and most importantly, release of mtDNA into cytosol. Previous studies demonstrated that mtDNA - a small 16569 bp long circular molecule present in multiple copies in mitochondria (Andrews *et al.*, 1999) - acts as a damage-associated molecular pattern (DAMP) in cytosol (Zhang *et al.*, 2010; Grazioli and Pugin, 2018). It regulates several key mechanisms of host immunity such as activation NF-κB (Zhang *et al.*, 2014) and NLRP3 inflammasome (Shimada *et al.*, 2012). Most importantly, our work has experimentally demonstrated that release of mtDNA upon fungal interaction leads to activation of interferon-stimulated genes (ISGs), which are the major players of host defence against various pathogens (Schneider, Chevillotte and Rice, 2014; Mills, Kelly and O'Neill, 2017). Moreover, we showed that the above mentioned mitochondrial events are not the typical signs of ongoing apoptosis, as it might be seen from the first sight. Instead, those are presumably the hallmarks of a recently discovered minority MOMP defence mechanism (Ichim *et al.*, 2015; Brokatzky *et al.*, 2019; Riley and Tait, 2020), when a limited number of mitochondria is permeabilized which instead of resulting in apoptosis, leads to partial activation of caspases and further triggering of cytokine expression. Minority MOMP, or sub-lethal mitochondrial signaling, has been demonstrated to be a defence mechanism against viruses and bacteria (Brokatzky *et al.*, 2019), and here we proposed that this conserved mechanism also acts against fungal pathogens.

Our findings open novel avenues for the research of epithelial responses against fungi and the role of the mitochondrial components in pathogen sensing and defence. For example, which specific parts of mtDNA act as a DAMP, how does mtDNA sequence variation affect the immune activation, how do mitochondria react to different fungal morphotypes, and many others. Moreover, considering that mitochondrial signaling leads to inflammation, which plays a central role in many diseases, including cancers, neurodegenerative and infectious diseases (Amor *et al.*, 2014; Chen *et al.*, 2018; Greten and Grivennikov, 2019), it potentially can be a target for novel therapeutics.

Cell damage as a driving force of human response

As discussed above, the initial response of epithelial cells to diverse *Candida* species was uniform. However, after 24hpi we observed a significant shift of human transcriptomic profile which was species-specific - *C. albicans* and *C. parapsilosis* induced cardinally different late human response, while it was similar in case of *C. glabrata* and *C. tropicalis*. We experimentally demonstrated that such a strong transcriptional shift directly depends on the extent of damage that each of *Candida* pathogens causes to humans. This was evident when we have infected human cells with *C. albicans ece1* deletion mutant, which elicited a similar response as *C. parapsilosis* - the least damaging species. This result indicates that damage is the major driving force of epithelial cell response against invading fungal pathogens. In fact, our observation serves as an direct experimental proof of the theoretical concept of “Damage-response framework of microbial pathogenicity” first suggested in the end of the last century (Casadevall and Pirofski, 1999) and further developed throughout recent decades (Casadevall and Pirofski, 2003, 2015). As the name states, this concept defines microbial pathogenicity as the net amount of host damage caused by the microbe throughout their interactions. Our results in essence can be seen as transcriptomic proof and ultimate positive control of this concept, since the immune system which protects the organism from damage is not present in our experimental model.

On the other hand, this concept is a broad generalization and thus does not imply any specific mechanism or type of damage. Interestingly, in the context of fungi, it seems that indeed the actual mechanism by which fungi cause damage supposedly does not play a critical role, since neither *C. glabrata* nor *C. tropicalis* express active candidalysin (in fact their damaging mechanisms are unknown). This is especially highlighted in the *C. glabrata* case, since it does not form true hyphae unlike *C. tropicalis*, which might (Jiang *et al.*, 2016) mechanically penetrate into host cells. Moreover, it was recently demonstrated that candidalysin alone is capable of triggering similar epithelial processes as *C. albicans*, such as damage, activation of pro-inflammatory cytokines and MAPK pathways (Richardson *et al.*, 2018). Thus, it might be of prime interest to disentangle whether those pathogenic effects are specific to candidalysin or are generic responses to any type of damage induced by different fungi, as it can be interpreted from the damage-response framework of microbial pathogenicity.

10.2 LncRNAs of major *Candida* pathogens

Chapter 5 continues the study of human-*Candida* interactions, but from a different angle - that of the fungal non-coding transcriptome.

Recent advances in high-throughput transcriptomics have shown that pervasive transcription is ubiquitous, and that the transcribed portion of eukaryotic genomes is much larger than previously anticipated (Tisseur, Kwapisz and Morillon, 2011; Hangauer, Vaughn and McManus, 2013; Kung, Colognori and Lee, 2013; Zhao *et al.*, 2016). While numerous types of non-coding RNAs have been identified in recent years, the lncRNAs have drawn a special interest among the scholars. Often called “A Dark Matter of Genomes”, lncRNAs - molecules longer than 200 bp which do not code for proteins - have been shown to play important roles in different fundamental biological processes, such as gene expression regulation, splicing, translation, imprinting, and cell cycle, among others (Merry, Niland and Khalil, 2015; Jandura and Krause, 2017; Zhang *et al.*, 2019).

Despite research of lncRNAs has skyrocketed in recent years, the majority of studies were carried out on model organisms such as humans or fruit fly (Iyer *et al.*, 2015; Wen *et al.*, 2016; Zhao *et al.*, 2016; Schor *et al.*, 2018; Uszczynska-Ratajczak *et al.*, 2018), and still there is only a handful of lncRNAs for which we understand the function at a mechanistic level. In this context, very little is known about these mysterious transcripts in *Candida* pathogens. Here, we catalogued and systematically characterized the lncRNAs of the main *Candida* pathogens, and took advantage of the comprehensive host-*Candida* interaction model described in Chapter 4 to assess their expression throughout the course of epithelial infection.

First, using complex and thorough bioinformatics analysis, we produced an exhaustive catalogues of lncRNAs of these species. Indeed, one could consider those catalogues exhaustive because we used all publicly available RNA-Seq datasets of those species (as of July 2019) in addition to dual RNA-Seq data of our host-*Candida* interaction model. In total we screened more than 2500 RNA-Seq datasets, encompassing many different biological conditions. The latter fact is particularly important considering that lncRNAs can be inducible under very specific conditions, and hence, our inclusive approach allowed us to capture lncRNA globally.

In accordance with studies on other fungi such as *S. cerevisiae* (David *et al.*, 2006; Kyriakou *et al.*, 2016), *Schizosaccharomyces pombe* (Atkinson *et al.*, 2018), *Neurospora crassa* (Arthanari *et al.*, 2014), *Coniophora puteana* (Borgognone *et al.*, 2019), *Metarhizium robertsii* (Z. Wang *et al.*, 2019) we found that long-non coding transcripts are very abundant in *Candida* species. For example, we observed that *C. albicans* has more lncRNAs (7216, mostly antisense transcripts) than protein-coding genes. The abundant lncRNA repertoires of the *Candida* species reinforces the idea that transcriptional capacity of fungal genomes is large (Niederer, Hass and Zappulla, 2017). On the other hand, the transcribability does not tacitly imply functional implications for these molecules, which ideally has to be backed up with experimental proves.

Compared to protein-coding genes, we found that lncRNAs of *Candida* species are shorter, have lower GC context, and have much lower expression levels. Interestingly, these three properties are common for lncRNA of a wide range of taxa, indicating that lncRNAs across the eukaryotic tree of life might have similar evolutionary constraints (Lopez-Ezquerra, Harrison and Bornberg-Bauer, 2017; Pegueroles, Iraola-Guzmán, *et al.*, 2019).

We observed poor sequence conservation, and higher synteny conservation between lncRNAs of the studied pathogens, with *C. glabrata* expectedly being an outlier among them. This kind of species-specificity is also observed in other taxa (Pegueroles, Iraola-Guzmán, *et al.*, 2019), which might suggest distinct roles of specific lncRNAs in different species. However, if species-specific lncRNAs might be interesting for further research, the small number of common transcripts observed between species undoubtedly deserves to be the focus of future studies.

Secondary structures of lncRNAs across species, as expected, were much more conserved compared to primary sequence and synteny. However, it still has to be further assessed whether structure conservation is a feasible parameter for assessing evolutionary relationships between lncRNAs in different species.

The co-expression analysis of lncRNAs and protein coding genes showed that lncRNAs are ubiquitous among highly co-expressed gene clusters, being in some cases the most interconnected nodes. Gene Ontology enrichment analysis showed a wide range of functional categories of these interconnected clusters, which, with caution, might be extrapolated to the lncRNAs involved in these clusters, serving as a first proxy for their functional roles. On the other hand, it has to be noted that co-expression analysis can report technically correct but biologically false-positive results, since unrelated transcripts also can be transcribed in a statistically co-expressed manner.

We took advantage of our model of epithelium-*Candida* interaction and assessed the expression dynamics of lncRNAs throughout the course of infection. For all fungal species, this analysis identified a large number of differentially expressed (DE) lncRNAs during interaction with human epithelial cells. Moreover, we identified infection-specific lncRNAs, which were exclusively DE during infection, but not in the culture medium. However, similarly to protein coding genes, most lncRNAs were DE in both conditions, implying that lncRNAs are as well involved in normal physiological activities of the yeasts. Despite the results of this analysis are largely descriptive, it serves as a powerful hypothesis-generating approach that identifies possible candidates for further experiments.

Overall, our study sheds light on the biology of these mysterious transcripts in the major *Candida* pathogens, and I believe, it will serve as an important starting point for further lncRNA research in fungal pathogens.

10.3 Enabling RNA-Seq for *in vivo* host-*Candida* interaction studies

In this PhD thesis we investigated the transcriptomic basis host-pathogen interactions between human and *Candida* species using an *in vitro* model of candidiasis. In fact, most of the other reports in this context also utilized various *in vitro* models (Tierney *et al.*, 2012; Rasheed, Battu and Kaur, 2018; Tóth *et al.*, 2018). This approach has its advantages, such as a controlled and adjustable experimental setup, low cost, relatively good reproducibility and less strict ethical regulations. On the other hand, a crucial drawback of *in vitro* models is that they might not fully resemble the mechanisms of host-pathogen interplay observed during the real infection (Fanning et al., 2012; Xu et al., 2015). However, a major problem precluding thorough transcriptomic analysis of host-pathogen interactions *in vivo* is an extremely low fungi/host ratio in the amounts of extracted RNA. Very low relative amounts of fungal RNA results in negligible amounts of fungal sequencing reads (less than 0.1%), allowing to analyze (if possible at all) only a handful of fungal genes (Liu *et al.*, 2015).

In Chapter 6, we report a methodology that allows to efficiently solve this problem by using a multiplexed targeted sequence enrichment based on SeqCap technology. We designed and tested oligonucleotide probes targeting the complete transcriptomes of the four major *Candida* species. We validated our probes using a large-scale dual transcriptomics analysis of human vaginal samples spiked with different loads of *C. albicans* cells.

Our enrichment design has several advantages over the only previously reported one (Amorim-Vaz *et al.*, 2015). First, we multiplexed the kit by designing it for enriching transcriptomes of four fungal species, while retaining the efficiency and accuracy of the enrichment. Considering that oligo kits even with minimal capacity are quite expensive, this kind of species multiplexing can be cost-effective when dealing with a small number of samples containing different species. Second, we designed the probes for enriching not only protein coding genes, but also for newly predicted lncRNAs (not the ones reported in Chapter 5). This would allow to analyze the transcriptional profiles of these molecules *in vivo*, which has never been reported for *Candida* pathogens. Finally, we show that RNA-Seq data obtained after targeted enrichment can be used for two additional types of analyses - fungal genotyping and host microbiome characterization. Both of these analyses are very valuable in the context of host-pathogen interaction. For example, one could correlate the microbiome composition of the host with gene expression profiles of the host or fungus at different stages of infection. Moreover, the fungal SNP data from clinical samples can be used to characterize fungal population structure or to link specific variants with resistance to antimycotic drugs. In general, these two layers of information opens a wide range of possible analyses.

It is worth mentioning that targeted enrichment with oligonucleotide probes is a relatively novel technique, especially for transcriptomics of host-pathogen interactions (Amorim-Vaz *et al.*, 2015; Chung *et al.*, 2018; Betin *et al.*, 2019). Thus, we still lack comprehensive benchmarking of different technologies compared to each other in this context. For example, SeqCap (Roche) and SureSelect (Agilent) are two alternative technologies (among others, like from Twist Biosciences, Qiagen, etc) which differ for each other by the configuration of probes - SeqCap uses overlapping probes, while SureSelect non-overlapping once. Thus, in my opinion, comprehensive benchmarking between these technologies at different setups and conditions is necessary to facilitate the choice between these technologies, depending on the specific needs.

10.4 Transcriptomic aftermath of hybridization in *C. orthopsilosis*

As previously hypothesized from phylogenetic inference (Gabaldón, Naranjo-Ortíz and Marcet-Houben, 2016) and demonstrated by comparative transcriptomics in this PhD project, the virulence of different *Candida* species is a result of a species-specific adaptation towards human host which has been developed several times independently across *Candida* lineage. Since it is unlikely that vertical evolution could have resulted in such phylogenetic distribution of pathogenic and non-pathogenic species, the plausible mechanism for this saltatory emergence of novel pathogens within short time frames was unclear. A recent series of studies has brought up a paradigm shift, showing that many *Candida* pathogens are hybrids (Pryszcz *et al.*, 2014, 2015; Schröder, Martinez de San Vicente, *et al.*, 2016; Mixão *et al.*, 2019) and that interspecies hybridization can potentially result in novel pathogenic species. Hybridization is an evolutionary powerful mechanism, which gives rise to hybrid organisms with novel adaptations that enable it to occupy new environmental niches, such as the human host (Gladieux *et al.*, 2014; Mixão and Gabaldón, 2018). Indeed, putative parental lineages - either one of both - of some hybrid yeast pathogens, such as *C. orthopsilosis* or *C. metapsilosis*, have never been observed among clinical isolates, which indicates a potential environmental lifestyle.

Hybridization impacts the whole biology of organisms, and thus far our understanding of the aftermaths of hybridization at different biological layers is still fragmented. Most research in this context has focused on a genomic level, where hybridization can result in genomic incompatibilities which are resolved through different mechanisms (such as gene conversion, large scale duplications and aneuploidies) leading to LOH (see Chapter 2). In contrast, the transcriptomic interactions between diverged regulatory networks of two parental species within a hybrid, and how these interactions might facilitate fungal virulence, are poorly understood.

Chapter 7 presents a project, in which we addressed these issues by investigating the allele-specific expression patterns in two independently formed strains of natural hybrid *C. orthopsilosis* and compared their gene expression levels with a likely parental strain. Importantly, the two hybrid strains analyzed in our study result from independent hybridizations of the same parentals but at different times - strain CP124 is thought to be a more recently formed hybrid (based on levels of LOH), while MCO454 is considered to be a more ancient hybrid. As a result, these two strains represent two extremes of heterozygosity levels, with CP124 being highly heterozygous, and MCO456 - highly homozoygous due to extensive LOH accumulated through time. Thus, our experimental design additionally enabled us to look into convergent evolution of these hybrids.

Allele-specific expression analysis with RNA-Seq of natural hybrids is technically challenging due to LOH, which eliminates the allele of one or another parent. The analysis of genes in heterozygosis showed common phased and ASE genes in two independently formed hybrids, indicating selective retention of some genes in heterozygous background and possible selection of allele-specific expression patterns across these strains.

The mechanism of this selection can be potentially explained by a hypothesis that LOH in these genes could be lethal for the hybrid. Thus, we only witness the strains of *C. orthopsilosis* that preferentially retained those genes in heterozygosis. A similar scenario could explain an allelic expression profile, which potentially might confer advantageous traits for the hybrid (though how it can be achieved mechanistically on the level of homeologous proteins and their functions is far from being clear). Unfortunately, the functions of the shared ASE genes in *C. orthopsilosis* are currently unknown to be able to confirm or reject these hypotheses. On the other hand, this opens new opportunities for further research. For example, on the genomic level it would be interesting to delete a copy of these genes in one and another parent, and trace the functional implications of these deletions. Alternatively, one could apply RNA interference to the transcripts of the homeologous genes, which could shed light on the selective advantages on the transcriptional level.

The comparisons between homozygous parental with the corresponding counterparts in the hybrids also showed moderate effect of hybridization. However, considering that the second parental species is still elusive, we do not know whether these differences in allelic expression are due to hybridization or they were already present in the parentals.

Importantly, in both types of comparisons - i.e. parent vs hybrid and homeologs vs each other, we observed DE/ASE genes which had virulence-related functions annotated in *C. albicans*, such as zinc and other metal ion metabolism. However, it is important to bear in mind that the functions of these genes are inferred from orthology relationships with *C. albicans* genes, which does not necessarily imply the exact functional recapitulation in this species. On the other hand, the majority of *C. orthopsilosis* genes are of unknown functions, and so even some hints from orthology and sequence conservation with other species can be useful.

Additionally, in our experimental setup we did not expose fungal strains to the host environment, but cultured them in YPD medium. However, as it was demonstrated previously (Liu *et al.*, 2015) and confirmed in this PhD project, transcriptional response of four main *Candida* species towards rich culture medium is similar to the response towards host cells, and it is unlikely that *C. orthopsilosis* would be an exception. Hence, one can expect to observe similar patterns of ASE and DE in this pathogen upon interaction with the host.

As research on fungal hybrids goes on, it is becoming increasingly obvious that hybridization has a significant contribution to the emergence of novel pathogens. In this context, comparative transcriptomics holds a great potential for unraveling the mechanistic principles of this contribution. To the best of our knowledge, our study represents the very first attempt to characterize the transcriptome interactions between parental genomes in a natural hybrid human fungal pathogen. Though we only scratched the surface of transcriptome complexity in these hybrids, several exciting findings were observed, such as moderate effect of hybridization on expression levels between parental and hybrids, presence of selection on heterozygosity and ASE, and imbalanced expression of genes possibly related to fungal virulence. Evidently, further research, involving other hybrid species, more biological conditions (especially the once resembling the host environment) and experimental testing, is necessary to conclusively answer the questions highlighted by our work.

10.5 Transcriptome shock is buffered in yeast hybrids

The transcriptomic analysis of the natural hybrid *C. orthopsilosis* (Chapter 7) revealed moderate changes in expression levels upon hybridization - ~4 % when comparing parent vs hybrid, and ~9% of ASE, which contrasts with what has been observed for animals and plants (McManus *et al.*, 2010; Wu *et al.*, 2018) However, considering that *C. orthopsilosis* is a natural evolved hybrid which has undergone LOH events to stabilize possible parental incompatibilities, these results can potentially be an underestimation of the real effect of hybridization on the transcriptomes of this species.

In the project described in Chapter 8 we aimed to disentangle the direct and immediate effect of hybridization on transcriptomes of yeast hybrids. To achieve this in an unbiased way, we had to analyze a hybrid where both parents are known and sequenced. Additionally, since we wanted to assess the immediate magnitude of transcriptome shock, a suitable model would have been a very recently formed hybrid. With these two criteria in mind, our choice has fallen on the artificially created hybrid of *S. cerevisiae* and *S. uvarum* - two distantly related species with ~20% of divergence in coding DNA (Kellis *et al.*, 2003). It has been previously hypothesized that genome shock should increase with an increase in genome divergence (Cox *et al.*, 2014), thus our model also can address this idea. To assess the impact of hybridization on the transcriptomes of these species, we did RNA-Seq of the hybrid and both parents at two terminal conditions, one of which (30 °C) was optimal for both species, while another (12 C°) favored only the cryotolerant *S. uvarum* (Li and Fay, 2017). This experimental design allowed us to compare the transcriptome differences between parents, between parents and hybrids, within hybrids and additionally compare the effect of transcriptome shock with the shock of environmental stress of temperature change.

Our approach allowed us to make several key observations. First, the magnitude of the transcriptional shock was surprisingly mild - hybridization altered 1-2% of genes in both species, which is in a stark contrast with the results observed in animals and plants. Moreover, the effect of temperature was stronger by ~6-fold. Second, there were more gene expression differences between orthologs of parental species than between homeologous genes within the hybrid. Finally, the results of ASE analysis showed that the majority of the imbalanced genes in the hybrid have inherited their expression levels from parentals.

Altogether, the results of these analyses suggest that, despite a substantial genome divergence between *S. cerevisiae* and *S. uvarum*, the transcriptional shock caused by hybridization is largely buffed in their hybrid.

We further aimed to clear up the reasons of the observed buffering and understand whether it is caused by active regulation of gene expression or is a passive outcome of independent co-existence of regulatory networks. Consistent with observation from gene expression data, we observed only minor changes in chromatin profiles and TF activity upon hybridization.

Thus, our integrative approach suggests little active regulatory cross-talk between the parental genomes of the studied hybrid (~91% of genes). The few changes in gene expression levels that we observed in this study can be explained by different scenarios of regulatory interplay. In one scenario parental orthologs equalize their expression levels after hybridization due to full compatibility of their regulatory mechanisms - a phenomenon called homeolog expression blending (Cox *et al.*, 2014), which accounts for ~6.5% of genes in our study. In an opposite scenario, homeologs can become differentially expressed after hybridization (~2.6% of genes), which potentially can be explained by selective targeting of one homeolog by the regulatory attributes of both homeologs. Despite the fact that these potential scenarios of interpreting the outcomes of hybridization are hypothetical, they can be experimentally tested using, for example, ChIP-Seq approach for detecting differential TF binding.

Overall, we observed a limited effect of hybridization on the transcriptome profiles of two distantly related *Saccharomyces* species. This observation is in agreement with our study of *C. orthopsilosis*, allowing us to argue that yeasts might possess an effective and universal (but likely a passive) way of dealing with transcriptional shock of hybridization. Whether this buffering is achieved by active regulatory interplays at some extent or represents a peaceful co-survival of independent transcriptomes, the existence of this buffering potentially can be at the root of the relatively high propensity of successful hybridization in Saccharomycotina yeasts. Further research with similar unbiased experimental setups involving other hybrids with varying genomic divergence and backed up with targeted follow-up experiments would allow to conclusively establish the mechanisms of the transcriptome buffering and its contribution to the emergence of novel, potentially pathogenic, fungal species.

10.6 Planning multi-species sequencing studies with Crossmapper

All studies of this PhD project were carried out with a methodological backbone of comparative transcriptomics. However, there is an additional technical nuance which is common across the projects - both host-pathogen interaction studies and research on hybrids deal with samples that contain genetic material from at least two different organisms. In the context of NGS, this peculiarity is not restricted to these areas of research, but in fact is common for microbiome analysis, metagenomics, and, generally speaking, for any system containing several species. Surprisingly and despite the research areas dealing with this problem are quite mainstream nowadays, there is no standard or even broadly accepted approach for separating the species sequencing data from each other. For host-pathogen interaction studies, for example, usually the data separation is achieved by mapping the mixture of data to reference genomes of analyzed species (Westermann, Barquist and Vogel, 2017). One major problem that arises with this approach is the possibility of cross-mapping - when reads from one species can map to another and vice-versa.

In this thesis, we propose a computational method, which does not solve the problem of cross-mapping *per se*, but allows us to avoid it in the first place. Chapter 9 presents a computational tool Crossmapper that allows designing multi-species sequencing projects so that the rates of cross-mapping are minimal.

As an additional valuable application, Crossmapper can also be used to save budgets for sequencing units or laboratories, which routinely sequence different species - DNA/RNA samples of different organisms can be mixed and sequenced as one sample, saving the cost for library preparation. Considering that the cost of sequencing is rapidly decreasing with an increase of throughput (Goodwin, McPherson and McCombie, 2016), the cost of library preparation can be the main bottleneck for large projects.

Thus, if the pooling of samples is planned in advance to avoid possible cross-mapping between species, sequencing entities can possibly save substantial financial resources on library preparations using Crossmapper.

Despite its great applicability potential, Crossmapper is a novel software relying on a simulation approach, and some of its applications ideally should be tested in advance, when possible. After all, even very precise simulations might not always reflect the real complexity and biases of sequencing procedures (Ross *et al.*, 2013; Boers, Jansen and Hays, 2019). For example, it would be useful to test the effect of mixed sequencing of several samples on some highly sensitive down-stream analyses, such as SNP calling or differential gene expression. Also, since the software uses the supplied genomes or transcriptomes for simulations, users need to feed Crossmapper with reference sequences which are as close as possible to the organisms of future study. Despite Crossmapper can simulate errors and mutations, the results of runs which use closely related species or strains (e.g. due to the absence of the actual references) need to be interpreted with caution.

From a technical perspective, Crossmapper still has a lot of room to grow. Its current implementation can be optimized to reduce memory consumption, run-times and storage requirements. Moreover, its major current applicability is project planning prior to sequencing, but in the future we can implement functions allowing to process the real data after sequencing, for example, by separating the mixture of reads to each corresponding genome. The list of useful features to implement will likely grow based on user feedback and our own ideas.

I hope that Crossmapper will serve as a useful tool for the research community. The area of its applicability is not just restricted to studies that were performed here, but goes far beyond. In my opinion, the problem of ambiguous read assignment between species has far deeper implications - the whole field of metagenomics and microbiome research is dealing with a similar obstacle, and though not being an expert in these fields, I think this issue needs to be systematically addressed and benchmarked, especially in the context of great expansion of sequencing capabilities.

10.7 Studying host-pathogen interactions in post-genomic era

There is no doubt that advances in next-generation sequencing have boosted biological research in the past two decades, and host-fungus interaction studies are no exception. Moreover, today when novel modifications of sequencing (e.g. single cell sequencing and long-read sequencing) are emerging, this technology becomes even more desired for many researchers. However, in my opinion, in the context of host-pathogen interactions, sequencing is not a Swiss army knife readily providing answers for posed questions. While for some disciplines the sequencing data itself readily holds the answer to a biological question of interest (i.e. evolutionary relationships between species), sequencing (especially transcriptome sequencing) mainly serves as a descriptive hypothesis-generating tool for host-pathogen interaction studies. After all, the major aim of a host-pathogen interplay study is to understand the mechanistic principles of how the organisms interact with each other in real time on a molecular level. For me personally it is hard to imagine a situation where transcriptome sequencing can readily answer that kind of mechanistic question without strong prior knowledge/background or subsequent experimentation. Moreover, when only a single gene out of many thousands can define the pathogenicity potential of a species (i.e. *ECE1* of *C. albicans*), finding this needle in a haystack can be very challenging.

The thoughts above are meant to deliver an idea that transcriptome sequencing can be necessary, but insufficient to understand the mechanisms of host-pathogen interactions. The same applies for deep understanding of virulence emergence. To reach these ultimate goals, further targeted experimentation needs to be carried out, which can require specific expertise from different fields, such as molecular biology, immunology, microbiology, microscopy, etc. A beneficial side of this issue is that bioinformatics personnel dealing with transcriptomics can be exposed to the principles and nuances of experimental biology and develop collaborative networks between peers.

While evidently the demand for transcriptome sequencing as a powerful approach for studying host-pathogen interplay will increase in the future, the potential users of this technology, in my opinion, should bear the above mentioned idea in mind, and be ready for a challenging high-risk-high-reward research.

Conclusions

Based on the results of this PhD thesis we have made the following conclusions:

- The four majors *Candida* pathogens have distinct species-specific transcriptomic responses towards human epithelium throughout the course of infection

- In contrast, human epithelial cells show a uniform response to the diverse *Candida* species at the early stage of infection, which is governed by a novel mitochondria-associated signaling pathway

- The response of human epithelium at the later stages of infection largely depends on the damaging capacity of each fungal pathogen

- The four major *Candida* pathogens encode numerous previously undescribed lncRNAs, which expression profiles hint to their potential implications in fungal virulence

- Targeted probe-based capturing is a highly effective method of enriching RNA of major *Candida* species from human-derived samples for further transcriptome sequencing and different downstream analyses

- Allele-specific expression in the fungal hybrid pathogen *C. orthopsilosis* is moderate, but at the same time shows signals of selection and potential links to pathogenic traits

- Hybridization in yeasts elicits a restricted transcriptomic shock, which is buffered due to a limited cross-talk between regulatory networks of parental species.

- The Crossmapper pipeline developed in this PhD project allows effortless planning of multi-species sequencing experiments

Appendix: List of publications

Hovhannisyan, H., Saus, E., Ksiezopolska, E., & Gabaldón, T. (2020). Integrative Omics Analysis Reveals a Limited Transcriptional Shock After Yeast Inter-species Hybridization. *Frontiers in Genetics*, *11*, 404.

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).

Hovhannisyan, H.*, Hafez, A.*, Llorens, C., & Gabaldón, T. (2020). CROSSMAPPER: estimating cross-mapping rates and optimizing experimental design in multi-species sequencing studies. *Bioinformatics*, *36*(3), 925-927.

Hovhannisyan, H., & Gabaldón, T. (2019). Transcriptome Sequencing Approaches to Elucidate Host–Microbe Interactions in Opportunistic Human Fungal Pathogens. *Current Topics of Immunology and Microbiology: Fungal Physiology and Immunopathogenesis*, 193-235.

OPATHY Consortium, Gabaldón, T. (2019). Recent trends in molecular diagnostics of yeast infections: from PCR to NGS. *FEMS microbiology reviews*, *43*(5), 517-547.

Pekmezovic M.*, Hovhannisyan H.*, Iracane E., Oliveira-Pacheco J., Seemann E., Qualmann B., Mogavero S., Gresnigt M., Brunke S., Butler G., Gabaldón T., Hube B. (2020). Diverse Candida pathogens induce protective mitochondria-associated type I interferon signalling and a damage-driven response in epithelial cells. *Nature Microbiology (under revision)*

Hovhannisyan H., Gabaldón, T. (2020). The lncRNA landscape of *Candida* pathogens. *(in preparation)*

Hovhannisyan H.*, Rodriguez A,*. Saus E., Vaneechoutte M., Gabaldón, T. Probe-based enrichment of fungal transcriptomes from human-derived samples enables *in vivo* study of *Candida* spp. infections *(in preparation).*

Hafez A., Hovhannisyan H., Molina M., Schikora-Tamarit MA, Llorens C., Gabaldón T. CandidaMine - an integrative database for omics data from *Candida* Species *(in preparation).*

Iraola-Guzmán S., Brunet A., Pegueroles C., Saus E., Hovhannisyan H., Casalots A., Pericay C., Gabaldón T. Zooming into the role of long non-coding RNAs in colorectal cancer by targeted transcriptional analysis of paraffin-embedded samples *(submitted).*

*equal contribution

References

Abad, A. *et al.* (2010) 'What makes Aspergillus fumigatus a successful pathogen? Genes and molecules involved in invasive aspergillosis', *Revista iberoamericana de micologia*, 27(4), pp. 155–182.

Adams, M. D. *et al.* (1991) 'Complementary DNA sequencing: expressed sequence tags and human genome project', *Science*, 252(5013), pp. 1651–1656.

Albertin, W. and Marullo, P. (2012) 'Polyploidy in fungi: evolution after whole-genome duplication', *Proceedings. Biological sciences / The Royal Society*, 279(1738), pp. 2497–2509.

Alwine, J. C., Kemp, D. J. and Stark, G. R. (1977) 'Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes', *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp. 5350–5354.

Amorim-Vaz, S. *et al.* (2015) 'RNA Enrichment Method for Quantitative Transcriptional Analysis of Pathogens In Vivo Applied to the Fungus Candida albicans', *mBio*, 6(5), pp. e00942–15.

Amorim-Vaz, S. and Sanglard, D. (2015) 'Novel Approaches for Fungal Transcriptomics from Host Samples', *Frontiers in microbiology*, 6, p. 1571.

Amor, S. *et al.* (2014) 'Inflammation in neurodegenerative diseases--an update', *Immunology*, 142(2), pp. 151–166.

Anders, S., Pyl, P. T. and Huber, W. (2015) 'HTSeq--a Python framework to work with high-throughput sequencing data', *Bioinformatics* , 31(2), pp. 166–169.

Andes, D. *et al.* (2005) 'A simple approach for estimating gene expression in Candida albicans directly from a systemic infection site', *The Journal of infectious diseases*, 192(5), pp. 893–900.

Andrews, R. M. *et al.* (1999) 'Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA', *Nature genetics*, 23(2), p. 147.

Aprianto, R. *et al.* (2016) 'Time-resolved dual RNA-seq reveals extensive rewiring of lung epithelial and pneumococcal transcriptomes during early infection’, *Genome biology*, 17(1), p. 198.

Arendrup, M. C. *et al.* (2014) ‘ESCMID and ECMM joint clinical guidelines for the diagnosis and management of rare invasive yeast infections’, *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 20 Suppl 3, pp. 76–98.

Arthanari, Y. *et al.* (2014) ‘Natural antisense transcripts and long non-coding RNA in Neurospora crassa’, *PloS one*, 9(3), p. e91353.

Atkinson, S. R. *et al.* (2018) ‘Long noncoding RNA repertoire and targeting by nuclear exosome, cytoplasmic exonuclease, and RNAi in fission yeast’, *RNA* , 24(9), pp. 1195–1213.

Au, K. F. *et al.* (2013) ‘Characterization of the human ESC transcriptome by hybrid sequencing’, *Proceedings of the National Academy of Sciences of the United States of America*, 110(50), pp. E4821–30.

Avital, G. *et al.* (2017) ‘scDual-Seq: mapping the gene regulatory program of Salmonella infection by host and pathogen single-cell RNA-sequencing’, *Genome biology*, 18(1), p. 200.

Avraham, R. *et al.* (2015) ‘Pathogen Cell-to-Cell Variability Drives Heterogeneity in Host Immune Responses’, *Cell*, 162(6), pp. 1309–1321.

Bahri, R. *et al.* (2010) ‘Normal human gingival epithelial cells sense C. parapsilosis by toll-like receptors and module its pathogenesis through antimicrobial peptides and proinflammatory cytokines’, *Mediators of inflammation*, 2010, p. 940383.

Bainbridge, M. N. *et al.* (2006) ‘Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach’, *BMC genomics*, 7, p. 246.

Baker, E. *et al.* (2015) ‘The Genome Sequence of Saccharomyces eubayanus and the Domestication of Lager-Brewing Yeasts’, *Molecular biology and evolution*, 32(11), pp. 2818–2831.

Baruzzo, G. *et al.* (2017) ‘Simulation-based comprehensive benchmarking of RNA-seq aligners’, *Nature methods*, 14(2), pp. 135–139.

Bateson, B. (no date) ‘Heredity and Variation in Modern Lights’, *William Bateson, Naturalist*, pp. 215–232. doi: 10.1017/cbo9780511693946.006.

Baxter, M. and Illston, G. M. (1980) ‘Temperature relationships of fungi isolated at low temperatures from soils and other substrates’, *Mycopathologia*, pp. 21–25. doi: 10.1007/bf00443047.

Berenguer, J. *et al.* (1993) ‘Lysis-centrifugation blood cultures in the detection of tissue-proven invasive candidiasis. Disseminated versus single-organ infection’, *Diagnostic microbiology and infectious disease*, 17(2), pp. 103–109.

Berman, J. (2016) ‘Ploidy plasticity: a rapid and reversible strategy for adaptation to stress’, *FEMS yeast research*, 16(3). doi: 10.1093/femsyr/fow020.

Betin, V. *et al.* (2019) ‘Hybridization-based capture of pathogen mRNA enables paired host-pathogen transcriptional analysis’, *Scientific reports*, 9(1), p. 19244.

Binkley, J. *et al.* (2014) ‘The Candida Genome Database: the new homology information page highlights protein similarity and phylogeny’, *Nucleic acids research*, 42(Database issue), pp. D711–6.

Bitar, D. *et al.* (2014) ‘Population-based analysis of invasive fungal infections, France, 2001-2010’, *Emerging infectious diseases*, 20(7), pp. 1149–1155.

Black, M. B. *et al.* (2014) ‘Comparison of microarrays and RNA-seq for gene expression analyses of dose-response experiments’, *Toxicological sciences: an official journal of the Society of Toxicology*, 137(2), pp. 385–403.

Blackwell, M. (2011) ‘The fungi: 1, 2, 3 ... 5.1 million species?’, *American journal of botany*, 98(3), pp. 426–438.

Boekhout, T. *et al.* (2001) ‘Hybrid genotypes in the pathogenic yeast Cryptococcus neoformans’, *Microbiology*, 147(Pt 4), pp. 891–907.

Boers, S. A., Jansen, R. and Hays, J. P. (2019) ‘Understanding and overcoming the pitfalls and biases of next-generation sequencing (NGS) methods for use in the routine clinical microbiological diagnostic laboratory’, *European journal of clinical microbiology & infectious diseases: official publication of the European Society of Clinical Microbiology*, 38(6), pp. 1059–1070.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) ‘Trimmomatic: a flexible trimmer for Illumina sequence data’, *Bioinformatics* , 30(15), pp. 2114–

2120.

Bongomin, F. *et al.* (2017) ‘Global and Multi-National Prevalence of Fungal Diseases—Estimate Precision’, *Journal of Fungi*, p. 57. doi: 10.3390/jof3040057.

Borgognone, A. *et al.* (2019) ‘Distribution, Characteristics, and Regulatory Potential of Long Noncoding RNAs in Brown-Rot Fungi’, *International journal of genomics and proteomics*, 2019, p. 9702342.

Borneman, A. R. and Pretorius, I. S. (2015) ‘Genomic insights into the Saccharomyces sensu stricto complex’, *Genetics*, 199(2), pp. 281–291.

Borodina, T., Adjaye, J. and Sultan, M. (2011) ‘A strand-specific library preparation protocol for RNA sequencing’, *Methods in enzymology*, 500, pp. 79–98.

Bovers, M. *et al.* (2008) ‘AIDS patient death caused by novel Cryptococcus neoformans x C. gattii hybrid’, *Emerging infectious diseases*, 14(7), pp. 1105–1108.

Brandão, F. *et al.* (2018) ‘HDAC genes play distinct and redundant roles in Cryptococcus neoformans virulence’, *Scientific reports*, 8(1), p. 5209.

Brandt, M. E. and Lockhart, S. R. (2012) ‘Recent Taxonomic Developments with Candida and Other Opportunistic Yeasts’, *Current Fungal Infection Reports*, pp. 170–177. doi: 10.1007/s12281-012-0094-x.

Bray, N. L. *et al.* (2016) ‘Near-optimal probabilistic RNA-seq quantification’, *Nature biotechnology*, 34(5), pp. 525–527.

Brokatzky, D. *et al.* (2019) ‘A non-death function of the mitochondrial apoptosis apparatus in immunity’, *The EMBO journal*, 38(11). doi: 10.15252/embj.2018100907.

Bronner, D. N. and O’Riordan, M. X. (2016) ‘Measurement of Mitochondrial DNA Release in Response to ER Stress’, *Bio-protocol*, 6(12). doi: 10.21769/BioProtoc.1839.

Brown, G. D. *et al.* (2012) ‘Hidden killers: human fungal infections’, *Science translational medicine*, 4(165), p. 165rv13.

Brown, N. A. *et al.* (2016) ‘RNAseq reveals hydrophobins that are involved in the adaptation of Aspergillus nidulans to lignocellulose’, *Biotechnology for biofuels*, 9, p. 145.

Broxton, C. N. and Culotta, V. C. (2016) ‘SOD Enzymes and Microbial Pathogens: Surviving the Oxidative Storm of Infection’, *PLoS pathogens*, 12(1), p. e1005295.

Bruno, V. M. *et al.* (2015) ‘Transcriptomic analysis of vulvovaginal candidiasis identifies a role for the NLRP3 inflammasome’, *mBio*, 6(2). doi: 10.1128/mBio.00182-15.

Buenrostro, J. D. *et al.* (2015) ‘ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide’, *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, 109, pp. 21.29.1–9.

Bullard, J. H. *et al.* (2010) ‘Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments’, *BMC bioinformatics*, 11, p. 94.

Burns, J. A. *et al.* (2017) ‘Transcriptome analysis illuminates the nature of the intracellular interaction in a vertebrate-algal symbiosis’. eLife Sciences Publications Limited. doi: 10.7554/eLife.22054.

Byrne, A. *et al.* (2017) ‘Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells’, *Nature communications*, 8, p. 16027.

Cabili, M. N. *et al.* (2011) ‘Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses’, *Genes & development*, 25(18), pp. 1915–1927.

Callaghan, S. and Guest, D. (2015) ‘Globalisation, the founder effect, hybrid Phytophthora species and rapid evolution: new headaches for biosecurity’, *Australasian Plant Pathology*, pp. 255–262. doi: 10.1007/s13313-015-0348-5.

Callahan, B. J. *et al.* (no date) ‘DADA2: High resolution sample inference from amplicon data’. doi: 10.1101/024034.

Campbell, J. D. *et al.* (2015) ‘Assessment of microRNA differential expression and detection in multiplexed small RNA sequencing data’, *RNA* , 21(2), pp. 164–171.

Campbell, M. A. *et al.* (2017) ‘Epichloë hybrida, sp. nov., an emerging model system for investigating fungal allopolyploidy’, *Mycologia*, 109(5), pp. 715–729.

Casadevall, A. and Pirofski, L. A. (1999) ‘Host-pathogen interactions:

redefining the basic concepts of virulence and pathogenicity', *Infection and immunity*, 67(8), pp. 3703–3713.

Casadevall, A. and Pirofski, L.-A. (2003) 'The damage-response framework of microbial pathogenesis', *Nature reviews. Microbiology*, 1(1), pp. 17–24.

Casadevall, A. and Pirofski, L.-A. (2015) 'What is a host? Incorporating the microbiota into the damage-response framework', *Infection and immunity*, 83(1), pp. 2–7.

Castel, S. E. *et al.* (2015) 'Tools and best practices for data processing in allelic expression analysis', *Genome biology*, 16, p. 195.

Cavalheiro, M. and Teixeira, M. C. (2018) 'Candida Biofilms: Threats, Challenges, and Promising Strategies', *Frontiers in Medicine*. doi: 10.3389/fmed.2018.00028.

Cemel, I. A. *et al.* (2017) 'The coding and noncoding transcriptome of Neurospora crassa', *BMC genomics*, 18(1), p. 978.

Cerqueira, G. C. *et al.* (2014) 'The Aspergillus Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations', *Nucleic acids research*, 42(Database issue), pp. D705–10.

Chacko, N. *et al.* (2015) 'The lncRNA RZE1 Controls Cryptococcal Morphological Transition', *PLoS genetics*, 11(11), p. e1005692.

Chalupová, J. *et al.* (2014) 'Identification of fungal microorganisms by MALDI-TOF mass spectrometry', *Biotechnology Advances*, pp. 230–241. doi: 10.1016/j.biotechadv.2013.11.002.

Chandra, J. *et al.* (2001) 'Biofilm formation by the fungal pathogen Candida albicans: development, architecture, and drug resistance', *Journal of bacteriology*, 183(18), pp. 5385–5394.

Chapman, B. *et al.* (2017) 'Changing epidemiology of candidaemia in Australia', *The Journal of antimicrobial chemotherapy*, 72(4), pp. 1103–1108.

Charlton, N. D. *et al.* (2014) 'Interspecific hybridization and bioactive alkaloid variation increases diversity in endophyticEpichloëspecies ofBromus laevipes', *FEMS Microbiology Ecology*, pp. 276–289. doi: 10.1111/1574-6941.12393.

Charron, G. *et al.* (2019) ‘Spontaneous whole-genome duplication restores fertility in interspecific hybrids’, *Nature communications*, 10(1), p. 4126.

Chen, F. *et al.* (2015) ‘Transcriptome Profiles of Human Lung Epithelial Cells A549 Interacting with Aspergillus fumigatus by RNA-Seq’, *PloS one*, 10(8), p. e0135720.

Chen, J. J. *et al.* (2007) ‘Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data’, *BMC Bioinformatics*. doi: 10.1186/1471-2105-8-412.

Chen, L. *et al.* (2014) ‘Transcriptional diversity during lineage commitment of human blood progenitors’, *Science*, 345(6204), p. 1251033.

Chen, L. *et al.* (2018) ‘Inflammatory responses and inflammation-associated diseases in organs’, *Oncotarget*, 9(6), pp. 7204–7218.

Chen, S.-Y. *et al.* (2017) ‘A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing’, *Scientific reports*, 7(1), p. 7648.

Chen, Y. *et al.* (2014) ‘The Cryptococcus neoformans transcriptome at the site of human meningitis’, *mBio*, 5(1), pp. e01087–13.

Cheon, S. A. *et al.* (2017) ‘A novel bZIP protein, Gsb1, is required for oxidative stress response, mating, and virulence in the human pathogen Cryptococcus neoformans’, *Scientific reports*, 7(1), p. 4044.

Cherry, J. M. *et al.* (2012) ‘Saccharomyces Genome Database: the genomics resource of budding yeast’, *Nucleic acids research*, 40(Database issue), pp. D700–5.

Chhangawala, S. *et al.* (2015) ‘The impact of read length on quantification of differentially expressed genes and splice junction detection’, *Genome biology*, 16, p. 131.

Choate, K. (2009) ‘Faculty Opinions recommendation of Combined immunodeficiency associated with DOCK8 mutations’, *Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature*. doi: 10.3410/f.1168490.630679.

Chowdhary, A., Sharma, C. and Meis, J. F. (2017) ‘Candida auris: A rapidly emerging cause of hospital-acquired multidrug-resistant fungal infections globally', *PLoS pathogens*, 13(5), p. e1006290.

Chu, E. Y. (2012) 'Cutaneous Manifestations of DOCK8 Deficiency Syndrome', *Archives of Dermatology*, p. 79. doi: 10.1001/archdermatol.2011.262.

Chung, M. *et al.* (2018) 'Targeted enrichment outperforms other enrichment techniques and enables more multi-species RNA-Seq analyses', *Scientific reports*, 8(1), p. 13377.

Cock, P. J. A. *et al.* (2010) 'The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants', *Nucleic acids research*, 38(6), pp. 1767–1771.

Combes, M.-C. *et al.* (2015) 'Regulatory divergence between parental alleles determines gene expression patterns in hybrids', *Genome biology and evolution*, 7(4), pp. 1110–1121.

Conesa, A. *et al.* (2016) 'A survey of best practices for RNA-seq data analysis', *Genome biology*, 17, p. 13.

Consortium OPATHY and Gabaldón, T. (2019) 'Recent trends in molecular diagnostics of yeast infections: from PCR to NGS', *FEMS microbiology reviews*, 43(5), pp. 517–547.

Cortegiani, A. *et al.* (2018) 'Epidemiology, clinical characteristics, resistance, and treatment of infections by Candida auris', *Journal of Intensive Care*. doi: 10.1186/s40560-018-0342-4.

Cottier, F. *et al.* (2015) 'The transcriptional stress response of Candida albicans to weak organic acids', *G3* , 5(4), pp. 497–505.

Cox, M. P. *et al.* (2014) 'An interspecific fungal hybrid reveals cross-kingdom rules for allopolyploid gene expression patterns', *PLoS genetics*, 10(3), p. e1004180.

Crawford, A. and Wilson, D. (2015) 'Essential metals at the host-pathogen interface: nutritional immunity and micronutrient assimilation by human fungal pathogens', *FEMS yeast research*, 15(7). doi: 10.1093/femsyr/fov071.

Croll, D., Zala, M. and McDonald, B. A. (2013) 'Breakage-fusion-bridge Cycles and Large Insertions Contribute to the Rapid Evolution of Accessory Chromosomes in a Fungal Pathogen', *PLoS Genetics*, p. e1003567. doi: 10.1371/journal.pgen.1003567.

Cubillos, F. A., Louis, E. J. and Liti, G. (2009) ‘Generation of a large set of genetically tractable haploid and diploid Saccharomyces strains’, *FEMS yeast research*, 9(8), pp. 1217–1225.

Dagenais, T. R. T. and Keller, N. P. (2009) ‘Pathogenesis of Aspergillus fumigatus in Invasive Aspergillosis’, *Clinical Microbiology Reviews*, pp. 447–465. doi: 10.1128/cmr.00055-08.

Dambuza, I. M. and Brown, G. D. (2018) ‘Sensing fungi at the oral epithelium’, *Nature microbiology*, pp. 4–5.

Danecek, P. *et al.* (2011) ‘The variant call format and VCFtools’, *Bioinformatics* , 27(15), pp. 2156–2158.

David, L. *et al.* (2006) ‘A high-resolution map of transcription in the yeast genome’, *Proceedings of the National Academy of Sciences of the United States of America*, 103(14), pp. 5320–5325.

del Fresno, C. *et al.* (2013) ‘Interferon-β Production via Dectin-1-Syk-IRF5 Signaling in Dendritic Cells Is Crucial for Immunity to C. albicans’, *Immunity*, pp. 1176–1186. doi: 10.1016/j.immuni.2013.05.010.

Depotter, J. R. *et al.* (2016) ‘Interspecific hybridization impacts host range and pathogenicity of filamentous microbes’, *Current opinion in microbiology*, 32, pp. 7–13.

Depotter, J. R. L. *et al.* (2016) ‘Verticillium longisporum, the invisible threat to oilseed rape and other brassicaceous plant hosts’, *Molecular plant pathology*, 17(7), pp. 1004–1016.

DePristo, M. A. *et al.* (2011) ‘A framework for variation discovery and genotyping using next-generation DNA sequencing data’, *Nature genetics*, 43(5), pp. 491–498.

Dijksterhuis, J., Houbraken, J. and Samson, R. A. (2013) ‘2 Fungal Spoilage of Crops and Food’, *Agricultural Applications*, pp. 35–56. doi: 10.1007/978-3-642-36821-9_2.

Dinel, S. *et al.* (2005) ‘Reproducibility, bioinformatic analysis and power of the SAGE method to evaluate changes in transcriptome’, *Nucleic acids research*, 33(3), p. e26.

Dobin, A. *et al.* (2013) ‘STAR: ultrafast universal RNA-seq aligner’, *Bioinformatics* , 29(1), pp. 15–21.

Dobin, A. and Gingeras, T. R. (2013) ‘Comment on TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions by Kim et al’, *Bioinformatics*. bioRxiv.

Dobzhansky, T. (1934) ‘Studies on hybrid sterility’, *Zeitschrift fur Zellforschung und Mikroskopische Anatomie,* pp. 169–223. doi: 10.1007/bf00374056.

Donaldson, M. E. *et al.* (2017) ‘Transcriptome analysis of smut fungi reveals widespread intergenic transcription and conserved antisense transcript expression’, *BMC genomics*, 18(1), p. 340.

D’Souza, C. A. *et al.* (2011) ‘Genome variation in Cryptococcus gattii, an emerging pathogen of immunocompetent hosts’, *mBio*, 2(1), pp. e00342–10.

Dujon, B. *et al.* (2004) ‘Genome evolution in yeasts’, *Nature*, 430(6995), pp. 35–44.

Dujon, B. (2010) ‘Yeast evolutionary genomics’, *Nature reviews. Genetics*, 11(7), pp. 512–524.

Dujon, B. A. and Louis, E. J. (2017) ‘Genome Diversity and Evolution in the Budding Yeasts (Saccharomycotina)’, *Genetics*, 206(2), pp. 717–750.

Dutta, A. *et al.* (2017) ‘Genome Dynamics of Hybrid Saccharomyces cerevisiae During Vegetative and Meiotic Divisions’, *G3: Genes|Genomes|Genetics*, pp. 3669–3679. doi: 10.1534/g3.117.1135.

Dutton, L. C. *et al.* (2016) ‘Transcriptional landscape of trans-kingdom communication between Candida albicans and Streptococcus gordonii’, *Molecular oral microbiology*, 31(2), pp. 136–161.

Ellepola, A. N. B. and Morrison, C. J. (2005) ‘Laboratory diagnosis of invasive candidiasis’, *Journal of microbiology* , 43 Spec No, pp. 65–84.

Ellstrand, N. C., Whitkus, R. and Rieseberg, L. H. (1996) ‘Distribution of spontaneous plant hybrids’, *Proceedings of the National Academy of Sciences of the United States of America*, 93(10), pp. 5090–5093.

Emmert-Streib, F. and Glazko, G. V. (2011) ‘Pathway analysis of expression data: deciphering functional building blocks of complex diseases’, *PLoS computational biology*, 7(5), p. e1002053.

Enguita, F. J. *et al.* (2016) ‘Transcriptomic Crosstalk between Fungal

Invasive Pathogens and Their Host Cells: Opportunities and Challenges for Next-Generation Sequencing Methods', *Journal of fungi (Basel, Switzerland)*, 2(1). doi: 10.3390/jof2010007.

Enjalbert, B., Nantel, A. and Whiteway, M. (2003) 'Stress-induced gene expression in Candida albicans: absence of a general stress response', *Molecular biology of the cell*, 14(4), pp. 1460–1467.

Ewels, P. *et al.* (2016) 'MultiQC: summarize analysis results for multiple tools and samples in a single report', *Bioinformatics*, pp. 3047–3048. doi: 10.1093/bioinformatics/btw354.

Fang, C., Wei, X. and Wei, Y. (2016) 'Mitochondrial DNA in the regulation of innate immune responses', *Protein & Cell*, pp. 11–16. doi: 10.1007/s13238-015-0222-9.

Fan, H. C., Fu, G. K. and Fodor, S. P. A. (2015) 'Expression profiling. Combinatorial labeling of single cells for gene expression cytometry', *Science*, 347(6222), p. 1258367.

Fanning, S. *et al.* (2012) 'Divergent targets of Candida albicans biofilm regulator Bcr1 in vitro and in vivo', *Eukaryotic cell*, 11(7), pp. 896–904.

Felk, A. *et al.* (2002) 'Candida albicans hyphal formation and the expression of the Efg1-regulated proteinases Sap4 to Sap6 are required for the invasion of parenchymal organs', *Infection and immunity*, 70(7), pp. 3689–3700.

Ferrareze, P. A. G. *et al.* (2017) 'Transcriptional Analysis Allows Genome Reannotation and Reveals that Cryptococcus gattii VGII Undergoes Nutrient Restriction during Infection', *Microorganisms*, 5(3). doi: 10.3390/microorganisms5030049.

Fidel, P. L. (2005) 'Immunity in vaginal candidiasis', *Current Opinion in Infectious Diseases*, pp. 107–111. doi: 10.1097/01.qco.0000160897.74492.a3.

Fidel, P. L., Jr *et al.* (2004) 'An intravaginal live Candida challenge in humans leads to new hypotheses for the immunopathogenesis of vulvovaginal candidiasis', *Infection and immunity*, 72(5), pp. 2939–2946.

Flevari, A. *et al.* (2013) 'Treatment of invasive candidiasis in the elderly: a review', *Clinical interventions in aging*, 8, pp. 1199–1208.

Francis, W. R. *et al.* (2013) 'A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly', *BMC Genomics*, p. 167. doi: 10.1186/1471-2164-14-167.

Fuller, K. K. *et al.* (2016) 'Aspergillus fumigatus Photobiology Illuminates the Marked Heterogeneity between Isolates', *mBio*, 7(5). doi: 10.1128/mBio.01517-16.

Gabaldón, T. *et al.* (2013) 'Comparative genomics of emerging pathogens in the Candida glabrata clade', *BMC genomics*, 14, p. 623.

Gabaldón, T. and Carreté, L. (2016) 'The birth of a deadly yeast: tracing the evolutionary emergence of virulence traits in Candida glabrata', *FEMS yeast research*, 16(2), p. fov110.

Gabaldón, T., Naranjo-Ortíz, M. A. and Marcet-Houben, M. (2016) 'Evolutionary genomics of yeast pathogens in the Saccharomycotina', *FEMS yeast research*, 16(6). doi: 10.1093/femsyr/fow064.

Galocha, M. *et al.* (2019) 'Divergent Approaches to Virulence in C. albicans and C. glabrata: Two Sides of the Same Coin', *International Journal of Molecular Sciences*, p. 2345. doi: 10.3390/ijms20092345.

Gao, Y. *et al.* (2019) 'Mitochondrial DNA Leakage Caused by Streptococcus pneumoniae Hydrogen Peroxide Promotes Type I IFN Expression in Lung Cells', *Frontiers in Microbiology*. doi: 10.3389/fmicb.2019.00630.

Garalde, D. R. *et al.* (2018) 'Highly parallel direct RNA sequencing on an array of nanopores', *Nature methods*, 15(3), pp. 201–206.

García-Sánchez, S. *et al.* (2004) 'Candida albicans biofilms: a developmental state associated with specific and stable gene expression patterns', *Eukaryotic cell*, 3(2), pp. 536–545.

Garcia-Solache, M. A. and Casadevall, A. (2010) 'Global warming will bring new fungal diseases for mammals', *mBio*, 1(1). doi: 10.1128/mBio.00061-10.

Geiss, G. K. *et al.* (2008) 'Direct multiplexed measurement of gene expression with color-coded probe pairs', *Nature biotechnology*, 26(3), pp. 317–325.

Gibbons, J. G. *et al.* (2012) 'Global transcriptome changes underlying colony growth in the opportunistic human pathogen Aspergillus fumigatus', *Eukaryotic cell*, 11(1), pp. 68–78.

Gillum, A. M., Tsay, E. Y. H. and Kirsch, D. R. (1984) ‘Isolation of the Candida albicans gene for orotidine-5′-phosphate decarboxylase by complementation of S. cerevisiae ura3 and E. coli pyrF mutations’, *Molecular and General Genetics MGG*, pp. 179–182. doi: 10.1007/bf00328721.

Gladieux, P. *et al.* (2011) ‘Maintenance of fungal pathogen species that are specialized to different hosts: allopatric divergence and introgression through secondary contact’, *Molecular biology and evolution*, 28(1), pp. 459–471.

Gladieux, P. *et al.* (2014) ‘Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes’, *Molecular ecology*, 23(4), pp. 753–773.

Goffeau, A. *et al.* (1996) ‘Life with 6000 genes’, *Science*, 274(5287), pp. 546, 563–7.

Gonzalez-Hilarion, S. *et al.* (2016) ‘Intron retention-dependent gene regulation in Cryptococcus neoformans’, *Scientific reports*, 6, p. 32252.

González-Torres, P. *et al.* (2015) ‘Interactions between closely related bacterial strains are revealed by deep transcriptome sequencing’, *Applied and environmental microbiology*, 81(24), pp. 8445–8456.

Goodwin, S., McPherson, J. D. and McCombie, W. R. (2016) ‘Coming of age: ten years of next-generation sequencing technologies’, *Nature reviews. Genetics*, 17(6), pp. 333–351.

Grant, P. R. and Grant, B. R. (1992) ‘Hybridization of bird species’, *Science*, 256(5054), pp. 193–197.

Graze, R. M. *et al.* (2012) ‘Allelic imbalance in Drosophila hybrid heads: exons, isoforms, and evolution’, *Molecular biology and evolution*, 29(6), pp. 1521–1532.

Grazioli, S. and Pugin, J. (2018) ‘Mitochondrial Damage-Associated Molecular Patterns: From Inflammatory Signaling to Human Diseases’, *Frontiers in immunology*, 9, p. 832.

Greaves, I. K., Gonzalez-Bayon, R., Wang, L., Zhu, A., Liu, P.-C., Groszmann, M., James Peacock, W., *et al.* (2015) ‘Epigenetic Changes in Hybrids’, *Plant Physiology*, pp. 1197–1205. doi: 10.1104/pp.15.00231.

Greaves, I. K., Gonzalez-Bayon, R., Wang, L., Zhu, A., Liu, P.-C., Groszmann, M., Peacock, W. J., *et al.* (2015) ‘Epigenetic Changes in Hybrids’, *Plant physiology*, 168(4), pp. 1197–1205.

Gresnigt, M. S. *et al.* (2012) ‘Neutrophil-mediated inhibition of proinflammatory cytokine responses’, *Journal of immunology* , 189(10), pp. 4806–4815.

Greten, F. R. and Grivennikov, S. I. (2019) ‘Inflammation and Cancer: Triggers, Mechanisms, and Consequences’, *Immunity*, 51(1), pp. 27–41.

Griffin, A. T. and Hanson, K. E. (2014) ‘Update on fungal diagnostics’, *Current infectious disease reports*, 16(8), p. 415.

Groszmann, M. *et al.* (2011) ‘Changes in 24-nt siRNA levels in Arabidopsis hybrids suggest an epigenetic contribution to hybrid vigor’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(6), pp. 2617–2622.

Gu, H. *et al.* (2019) ‘Inheritance patterns of the transcriptome in hybrid chickens and their parents revealed by expression analysis’, *Scientific reports*, 9(1), p. 5750.

Guinea, J. (2014) ‘Global trends in the distribution of Candida species causing candidemia’, *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 20 Suppl 6, pp. 5–10.

Guo, B. *et al.* (2013) ‘Comparative proteomic analysis of embryos between a maize hybrid and its parental lines during early stages of seed germination’, *PloS one*, 8(6), p. e65867.

Guo, M. *et al.* (2008) ‘Genome-wide allele-specific expression analysis using Massively Parallel Signature Sequencing (MPSSTM) Reveals cis- and trans-effects on gene expression in maize hybrid meristem tissue’, *Plant Molecular Biology*, pp. 551–563. doi: 10.1007/s11103-008-9290-z.

Guo, X. *et al.* (2016) ‘Advances in long noncoding RNAs: identification, structure prediction and function annotation’, *Briefings in functional genomics*, 15(1), pp. 38–46.

Guo, Y. *et al.* (2014) ‘RNAseqPS: A Web Tool for Estimating Sample Size and Power for RNAseq Experiment’, *Cancer informatics*, 13(Suppl 6), pp. 1–5.

Gusmao, E. G. *et al.* (2014) ‘Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications', *Bioinformatics* , 30(22), pp. 3143–3151.

Haas, B. J. *et al.* (2013) 'De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis', *Nature protocols*, 8(8), pp. 1494–1512.

Hagen, F. *et al.* (2015) 'Recognition of seven species in the Cryptococcus gattii/Cryptococcus neoformans species complex', *Fungal genetics and biology: FG & B*, 78, pp. 16–48.

Hahn, S. and Young, E. T. (2011) 'Transcriptional regulation in Saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators', *Genetics*, 189(3), pp. 705–736.

Hangauer, M. J., Vaughn, I. W. and McManus, M. T. (2013) 'Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs', *PLoS genetics*, 9(6), p. e1003569.

Hart, S. N. *et al.* (2013) 'Calculating sample size estimates for RNA sequencing data', *Journal of computational biology: a journal of computational molecular cell biology*, 20(12), pp. 970–978.

Havlickova, B., Czaika, V. A. and Friedrich, M. (2008) 'Epidemiological trends in skin mycoses worldwide', *Mycoses*, 51 Suppl 4, pp. 2–15.

Hawksworth, D. L. and Lücking, R. (2017) 'Fungal Diversity Revisited: 2.2 to 3.8 Million Species', *Microbiology spectrum*, 5(4). doi: 10.1128/microbiolspec.FUNK-0052-2016.

Hebecker, B. *et al.* (2016) 'Corrigendum: Dual-species transcriptional profiling during systemic candidiasis reveals organ-specific host-pathogen interactions', *Scientific reports*, 6, p. 39423.

Heil, C. S. S. *et al.* (2017) 'Loss of Heterozygosity Drives Adaptation in Hybrid Yeast', *Molecular Biology and Evolution*, pp. 1596–1612. doi: 10.1093/molbev/msx098.

Hernandez, R. and Rupp, S. (2009) 'Human Epithelial Model Systems for the Study of Candida infections In Vitro: Part II. Histologic Methods for Studying Fungal Invasion', *Host-Pathogen Interactions*, pp. 105–123. doi: 10.1007/978-1-59745-204-5_10.

Heward, J. A. and Lindsay, M. A. (2014) ‘Long non-coding RNAs in the regulation of the immune response’, *Trends in Immunology*, pp. 408–419. doi: 10.1016/j.it.2014.07.005.

Hovhannisyan, H., Hafez, A., *et al.* (2020) ‘CROSSMAPPER: estimating cross-mapping rates and optimizing experimental design in multi-species sequencing studies’, *Bioinformatics* , 36(3), pp. 925–927.

Hovhannisyan, H., Saus, E., *et al.* (2020) ‘Integrative Omics Analysis Reveals a Limited Transcriptional Shock After Yeast Interspecies Hybridization’, *Frontiers in Genetics*. doi: 10.3389/fgene.2020.00404.

Hovhannisyan, H. and Gabaldón, T. (2019) ‘Transcriptome Sequencing Approaches to Elucidate Host-Microbe Interactions in Opportunistic Human Fungal Pathogens’, *Current topics in microbiology and immunology*, 422, pp. 193–235.

Hoyer, L. L. and Cota, E. (2016) ‘Candida albicans Agglutinin-Like Sequence (Als) Family Vignettes: A Review of Als Protein Structure and Function’, *Frontiers in microbiology*, 7, p. 280.

Hu, B. *et al.* (2011) ‘Pathogen comparative genomics in the next-generation sequencing era: genome alignments, pangenomics and metagenomics’, *Briefings in functional genomics*, 10(6), pp. 322–333.

Hu, G. *et al.* (2014) ‘Microevolution during serial mouse passage demonstrates FRE3 as a virulence adaptation gene in Cryptococcus neoformans’, *mBio*, 5(2), pp. e00941–14.

Hunt, S. E. *et al.* (2018) ‘Ensembl variation resources’, *Database: the journal of biological databases and curation*, 2018. doi: 10.1093/database/bay119.

Hu, X. *et al.* (2016) ‘Transcriptome profiling and comparison of maize ear heterosis during the spikelet and floret differentiation stages’, *BMC genomics*, 17(1), p. 959.

Hu, X. *et al.* (2017) ‘Genome-wide proteomic profiling reveals the role of dominance protein expression in heterosis in immature maize ears’, *Scientific reports*, 7(1), p. 16130.

Ichim, G. *et al.* (2015) ‘Limited mitochondrial permeabilization causes DNA damage and genomic instability in the absence of cell death’, *Molecular cell*, 57(5), pp. 860–872.

Idnurm, A. *et al.* (2009) ‘Identification of ENA1 as a virulence gene of the human pathogenic fungus Cryptococcus neoformans through signature-tagged insertional mutagenesis’, *Eukaryotic cell*, 8(3), pp. 315–326.

Irmer, H. *et al.* (2015) ‘RNAseq analysis of Aspergillus fumigatus in blood reveals a just wait and see resting stage behavior’, *BMC genomics*, 16, p. 640.

Iyer, M. K. *et al.* (2015) ‘The landscape of long noncoding RNAs in the human transcriptome’, *Nature genetics*, 47(3), pp. 199–208.

Jabra-Rizk, M. A. *et al.* (2016) ‘Candida albicans Pathogenesis: Fitting within the Host-Microbe Damage Response Framework’, *Infection and immunity*, 84(10), pp. 2724–2739.

Jaeger, M. *et al.* (2015) ‘The RIG-I-like helicase receptor MDA5 (IFIH1) is involved in the host defense against Candida infections’, *European Journal of Clinical Microbiology & Infectious Diseases*, pp. 963–974. doi: 10.1007/s10096-014-2309-2.

Jain, M. *et al.* (2018) ‘Nanopore sequencing and assembly of a human genome with ultra-long reads’, *Nature biotechnology*, 36(4), pp. 338–345.

Jamuar, S. S. and Tan, E.-C. (2015) ‘Clinical application of next-generation sequencing for Mendelian diseases’, *Human genomics*, 9, p. 10.

Janbon, G. *et al.* (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), p. e1004261.

Jandura, A. and Krause, H. M. (2017) ‘The New RNA World: Growing Evidence for Long Noncoding RNA Functionality’, *Trends in Genetics*, pp. 665–676. doi: 10.1016/j.tig.2017.08.002.

Jarroux, J., Morillon, A. and Pinskaya, M. (2017) ‘History, Discovery, and Classification of lncRNAs’, *Advances in experimental medicine and biology*, 1008, pp. 1–46.

Jiang, C. *et al.* (2016) ‘Significance of hyphae formation in virulence of Candida tropicalis and transcriptomic analysis of hyphal cells’, *Microbiological research*, 192, pp. 65–72.

Jiang, H. *et al.* (2014) ‘Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads', *BMC bioinformatics*, 15, p. 182.

Jiang, M. *et al.* (2018) 'Self-Recognition of an Inducible Host lncRNA by RIG-I Feedback Restricts Innate Immune Response', *Cell*, 173(4), pp. 906–919.e13.

Jia, X. *et al.* (2014) 'Gliotoxin promotes Aspergillus fumigatus internalization into type II human pneumocyte A549 cells by inducing host phospholipase D activation', *Microbes and infection / Institut Pasteur*, 16(6), pp. 491–501.

Jimenez, A., Tipper, D. J. and Davies, J. (1973) 'Mode of action of thiolutin, an inhibitor of macromolecular synthesis in Saccharomyces cerevisiae', *Antimicrobial agents and chemotherapy*, 3(6), pp. 729–738.

Johnsson, P. *et al.* (2014) 'Evolutionary conservation of long non-coding RNAs; sequence, structure, function', *Biochimica et biophysica acta*, 1840(3), pp. 1063–1071.

Jones, A. N. and Sattler, M. (2019) 'Challenges and perspectives for structural biology of lncRNAs—the example of the Xist lncRNA A-repeats', *Journal of Molecular Cell Biology*, pp. 845–859. doi: 10.1093/jmcb/mjz086.

Jonge, R. de *et al.* (2013) 'Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen', *Genome Research*, pp. 1271–1282. doi: 10.1101/gr.152660.112.

Jung, W. H. *et al.* (2009) 'Role of ferroxidases in iron uptake and virulence of Cryptococcus neoformans', *Eukaryotic cell*, 8(10), pp. 1511–1520.

Jung, W. H. (2015) 'The Zinc Transport Systems and Their Regulation in Pathogenic Fungi', *Mycobiology*, 43(3), pp. 179–183.

Käding, N. *et al.* (2017) 'Growth of Chlamydia pneumoniae Is Enhanced in Cells with Impaired Mitochondrial Function', *Frontiers in Cellular and Infection Microbiology*. doi: 10.3389/fcimb.2017.00499.

Kale, S. D. *et al.* (2017) 'Modulation of Immune Signaling and Metabolism Highlights Host and Fungal Transcriptional Responses in Mouse Models of Invasive Pulmonary Aspergillosis', *Scientific reports*, 7(1), p. 17096.

Kathiravan, M. K. *et al.* (2012) ‘The biology and chemistry of antifungal agents: a review’, *Bioorganic & medicinal chemistry*, 20(19), pp. 5678–5698.

Kearney, C. J., Randall, K. L. and Oliaro, J. (2017) ‘DOCK8 regulates signal transduction events to control immunity’, *Cellular & molecular immunology*, 14(5), pp. 406–411.

Kebaara, B. W. *et al.* (2006) ‘Determination of mRNA half-lives in Candida albicans using thiolutin as a transcription inhibitor’, *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada*, 49(8), pp. 894–899.

Kellis, M. *et al.* (2003) ‘Sequencing and comparison of yeast species to identify genes and regulatory elements’, *Nature*, 423(6937), pp. 241–254.

Khan, A. *et al.* (2018) ‘JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework’, *Nucleic acids research*, 46(D1), p. D1284.

Khot, P. D. and Fredricks, D. N. (2009) ‘PCR-based diagnosis of human fungal infections’, *Expert Review of Anti-infective Therapy*, pp. 1201–1221. doi: 10.1586/eri.09.104.

Kim, D. *et al.* (2013) ‘TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions’, *Genome biology*, 14(4), p. R36.

Kim, D. *et al.* (2016) ‘Centrifuge: rapid and sensitive classification of metagenomic sequences’, *Genome research*, 26(12), pp. 1721–1729.

Kim, D., Langmead, B. and Salzberg, S. L. (2015) ‘HISAT: a fast spliced aligner with low memory requirements’, *Nature methods*, 12(4), pp. 357–360.

Kim, E. S. *et al.* (2014) ‘Mitochondrial dynamics regulate melanogenesis through proteasomal degradation of MITF via ROS-ERK activation’, *Pigment cell & melanoma research*, 27(6), pp. 1051–1062.

Kim, J. and Sudbery, P. (2011) ‘Candida albicans, a major human fungal pathogen’, *The Journal of Microbiology*, pp. 171–177. doi: 10.1007/s12275-011-1064-7.

Kim, M. *et al.* (2010) ‘Bacterial Interactions with the Host Epithelium’, *Cell Host & Microbe*, pp. 20–35. doi: 10.1016/j.chom.2010.06.006.

Kim, W. *et al.* (2018) ‘Developmental Dynamics of Long Noncoding RNA Expression during Sexual Fruiting Body Formation in Fusarium graminearum’, *mBio*, 9(4). doi: 10.1128/mBio.01292-18.

Klingspor, L. *et al.* (2015) ‘Invasive Candida infections in surgical patients in intensive care units: a prospective, multicentre survey initiated by the European Confederation of Medical Mycology (ECMM) (2006-2008)’, *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 21(1), pp. 87.e1–87.e10.

Kolisko, M. *et al.* (2014) ‘Single-cell transcriptomics for microbial eukaryotes’, *Current biology: CB*, 24(22), pp. R1081–2.

Kolodziejczyk, A. A. *et al.* (2015) ‘The Technology and Biology of Single-Cell RNA Sequencing’, *Molecular Cell*, pp. 610–620. doi: 10.1016/j.molcel.2015.04.005.

Kong, L. *et al.* (2007) ‘CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine’, *Nucleic acids research*, 35(Web Server issue), pp. W345–9.

Kotredes, K. P., Thomas, B. and Gamero, A. M. (2017) ‘The Protective Role of Type I Interferons in the Gastrointestinal Tract’, *Frontiers in immunology*, 8, p. 410.

Kowalski, C. H. *et al.* (2016) ‘Heterogeneity among Isolates Reveals that Fitness in Low Oxygen Correlates with Aspergillus fumigatus Virulence’, *mBio*, 7(5). doi: 10.1128/mBio.01515-16.

Kozel, T. R. and Wickes, B. (2014) ‘Fungal diagnostics’, *Cold Spring Harbor perspectives in medicine*, 4(4), p. a019299.

Krogerus, K. *et al.* (2017) ‘Novel brewing yeast hybrids: creation and application’, *Applied microbiology and biotechnology*, 101(1), pp. 65–78.

Ksiezopolska, E. and Gabaldón, T. (2018) ‘Evolutionary Emergence of Drug Resistance in Candida Opportunistic Pathogens’, *Genes*, 9(9). doi: 10.3390/genes9090461.

Kung, J. T. Y., Colognori, D. and Lee, J. T. (2013) ‘Long noncoding RNAs: past, present, and future’, *Genetics*, 193(3), pp. 651–669.

Kupfer, D. M. *et al.* (2004) ‘Introns and splicing elements of five diverse fungi’, *Eukaryotic cell*, 3(5), pp. 1088–1100.

Kurihara, Y. *et al.* (2019) ‘Chlamydia trachomatis targets mitochondrial dynamics to promote intracellular survival and proliferation’, *Cellular microbiology*, 21(1), p. e12962.

Kutter, C. *et al.* (2012) ‘Rapid turnover of long noncoding RNAs and the evolution of gene expression’, *PLoS genetics*, 8(7), p. e1002841.

Kwon-Chung, K. J. *et al.* (2014) ‘Cryptococcus neoformans and Cryptococcus gattii, the etiologic agents of cryptococcosis’, *Cold Spring Harbor perspectives in medicine*, 4(7), p. a019760.

Kwon-Chung, K. J. and Sugui, J. A. (2013) ‘Aspergillus fumigatus--what makes the species a ubiquitous human fungal pathogen?’, *PLoS pathogens*, 9(12), p. e1003743.

Kyriakou, D. *et al.* (2016) ‘Functional characterisation of long intergenic non-coding RNAs through genetic interaction profiling in Saccharomyces cerevisiae’, *BMC biology*, 14(1), p. 106.

Langfelder, P. and Horvath, S. (2008) ‘WGCNA: an R package for weighted correlation network analysis’, *BMC bioinformatics*. BioMed Central, 9(1), pp. 1–13.

Latgé, J. P. (1999) ‘Aspergillus fumigatus and aspergillosis’, *Clinical microbiology reviews*, 12(2), pp. 310–350.

Leinonen, R. *et al.* (2011) ‘The sequence read archive’, *Nucleic acids research*, 39(Database issue), pp. D19–21.

Le Jeune, C. *et al.* (2007) ‘Characterization of natural hybrids of Saccharomyces cerevisiae and Saccharomyces bayanus var. uvarum’, *FEMS yeast research*, 7(4), pp. 540–549.

Lelli, K. M., Slattery, M. and Mann, R. S. (2012) ‘Disentangling the many layers of eukaryotic transcriptional regulation’, *Annual review of genetics*, 46, pp. 43–68.

Levin, J. Z. *et al.* (2010) ‘Comprehensive comparative analysis of strand-specific RNA sequencing methods’, *Nature methods*, 7(9), pp. 709–715.

Li, A. *et al.* (2014) ‘mRNA and Small RNA Transcriptomes Reveal Insights into Dynamic Homoeolog Regulation of Allopolyploid Heterosis in Nascent Hexaploid Wheat’, *The Plant cell*, 26(5), pp. 1878–1900.

Liao, Y., Smyth, G. K. and Shi, W. (2014) ‘featureCounts: an efficient general purpose program for assigning sequence reads to genomic features', *Bioinformatics* , 30(7), pp. 923–930.

Li, C. X. *et al.* (2015) 'Candida albicans adapts to host copper during infection by swapping metal cofactors for superoxide dismutase', *Proceedings of the National Academy of Sciences of the United States of America*, 112(38), pp. E5336–42.

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

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows-Wheeler transform', *Bioinformatics* , 25(14), pp. 1754–1760.

Li, J. *et al.* (2014) 'Genome-wide identification and characterization of long intergenic non-coding RNAs in Ganoderma lucidum', *PloS one*, 9(6), p. e99442.

Lin, J.-Q. *et al.* (2013) 'Transcriptomic profiling of Aspergillus flavus in response to 5-azacytidine', *Fungal genetics and biology: FG & B*, 56, pp. 78–86.

Lionakis, M. S., Iliev, I. D. and Hohl, T. M. (2017) 'Immunity against fungi', *JCI insight*, 2(11). doi: 10.1172/jci.insight.93156.

Lister, R. *et al.* (2008) 'Highly integrated single-base resolution maps of the epigenome in Arabidopsis', *Cell*, 133(3), pp. 523–536.

Li, T. *et al.* (2019) 'Therapeutic effectiveness of type I interferon in vulvovaginal candidiasis', *Microbial pathogenesis*, 134, p. 103562.

Liti, G. *et al.* (2009) 'Population genomics of domestic and wild yeasts', *Nature*, 458(7236), pp. 337–341.

Liu, H. *et al.* (2018) 'Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias', *Nature ecology & evolution*, 2(1), pp. 164–173.

Liu, T.-B. *et al.* (2014) 'Cryptococcus inositol utilization modulates the host protective immune response during brain infection', *Cell communication and signaling: CCS*, 12, p. 51.

Liu, Y. *et al.* (2013) 'Evaluating the impact of sequencing depth on transcriptome profiling in human adipose', *PloS one*, 8(6), p. e66883.

Liu, Y. *et al.* (2015) ‘New signaling pathways govern the host response to C. albicans infection in various niches’, *Genome research*, 25(5), pp. 679–689.

Liu, Y. and Filler, S. G. (2011) ‘Candida albicans Als3, a multifunctional adhesin and invasin’, *Eukaryotic cell*, 10(2), pp. 168–173.

Liu, Y., Zhou, J. and White, K. P. (2014) ‘RNA-seq differential expression studies: more sequence or more replication?’, *Bioinformatics* , 30(3), pp. 301–304.

Li, X. C. and Fay, J. C. (2017) ‘Cis-Regulatory Divergence in Gene Expression between Two Thermally Divergent Yeast Species’, *Genome biology and evolution*, 9(5), pp. 1120–1129.

Li, Z. *et al.* (2019) ‘Identification of transcription factor binding sites using ATAC-seq’, *Genome biology*, 20(1), p. 45.

Lockhart, D. J. *et al.* (1996) ‘Expression monitoring by hybridization to high-density oligonucleotide arrays’, *Nature biotechnology*, 14(13), pp. 1675–1680.

Lopez-Ezquerra, A., Harrison, M. C. and Bornberg-Bauer, E. (2017) ‘Comparative analysis of lincRNA in insect species’, *BMC evolutionary biology*, 17(1), p. 155.

Lopez-Maestre, H. *et al.* (2017) ‘Identification of misexpressed genetic elements in hybrids between Drosophila-related species’, *Scientific reports*, 7, p. 40618.

Lorenz, M. C., Bender, J. A. and Fink, G. R. (2004) ‘Transcriptional response of Candida albicans upon internalization by macrophages’, *Eukaryotic cell*, 3(5), pp. 1076–1087.

Lorenz, R. *et al.* (2011) ‘ViennaRNA Package 2.0’, *Algorithms for molecular biology: AMB*, 6, p. 26.

Losada, L. *et al.* (2014) ‘Large-scale transcriptional response to hypoxia in Aspergillus fumigatus observed using RNAseq identifies a novel hypoxia regulated ncRNA’, *Mycopathologia*, 178(5-6), pp. 331–339.

Love, M. I., Huber, W. and Anders, S. (2014) ‘Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2’, *Genome biology*, 15(12), p. 550.

Lowe, R. *et al.* (2017) 'Transcriptomics technologies', *PLoS computational biology*, 13(5), p. e1005457.

Lu, H., Giordano, F. and Ning, Z. (2016) 'Oxford Nanopore MinION Sequencing and Genome Assembly', *Genomics, proteomics & bioinformatics*, 14(5), pp. 265–279.

Luthra, R. *et al.* (2015) 'Next-Generation Sequencing in Clinical Molecular Diagnostics of Cancer: Advantages and Challenges', *Cancers*, 7(4), pp. 2023–2036.

MacCallum, D. M. (2012) 'Hosting Infection: Experimental Models to AssayCandidaVirulence', *International Journal of Microbiology*, pp. 1–12. doi: 10.1155/2012/363764.

Maguire, S. L. *et al.* (2013) 'Comparative Genome Analysis and Gene Finding in Candida Species Using CGOB', *Molecular Biology and Evolution*, pp. 1281–1291. doi: 10.1093/molbev/mst042.

Makanjuola, O., Bongomin, F. and Fayemiwo, S. (2018) 'An Update on the Roles of Non-albicans Candida Species in Vulvovaginitis', *Journal of Fungi*, p. 121. doi: 10.3390/jof4040121.

Mallet, J. (2005) 'Hybridization as an invasion of the genome', *Trends in Ecology & Evolution*, pp. 229–237. doi: 10.1016/j.tree.2005.02.010.

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

Mårdh, P.-A. *et al.* (2002) 'Facts and myths on recurrent vulvovaginal candidosis—a review on epidemiology, clinical manifestations, diagnosis, pathogenesis and therapy', *International Journal of STD & AIDS*, pp. 522–539. doi: 10.1258/095646202760159639.

Marioni, J. C. *et al.* (2008) 'RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays', *Genome research*, 18(9), pp. 1509–1517.

Marquez, L. M. *et al.* (2007) 'A Virus in a Fungus in a Plant: Three-Way Symbiosis Required for Thermal Tolerance', *Science*, pp. 513–515. doi: 10.1126/science.1136237.

Martchenko, M. *et al.* (2004) 'Superoxide dismutases in Candida albicans:

transcriptional regulation and functional characterization of the hyphal-induced SOD5 gene', *Molecular biology of the cell*, 15(2), pp. 456–467.

Martin, R. *et al.* (2013) 'A core filamentation response network in Candida albicans is restricted to eight genes', *PloS one*, 8(3), p. e58613.

Mattei, E. *et al.* (2015) 'Web-Beagle: a web server for the alignment of RNA secondary structures', *Nucleic acids research*, 43(W1), pp. W493–7.

Mayer, F. L., Wilson, D. and Hube, B. (2013) 'Candida albicans pathogenicity mechanisms', *Virulence*, 4(2), pp. 119–128.

May, R. C. *et al.* (2016) 'Cryptococcus: from environmental saprophyte to global pathogen', *Nature reviews. Microbiology*, 14(2), pp. 106–117.

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

McCoy, R. C. *et al.* (2014) 'Illumina TruSeq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements', *PloS one*, 9(9), p. e106689.

McGhee, S. A. *et al.* (2010) 'DOCK8 Deletions and Mutations Are Associated With The Autosomal Recessive Hyper-IgE Phenotype', *Journal of Allergy and Clinical Immunology*, p. AB356. doi: 10.1016/j.jaci.2010.01.006.

McManus, C. J. *et al.* (2010) 'Regulatory divergence in Drosophila revealed by mRNA-seq', *Genome research*, 20(6), pp. 816–825.

Meir, J. *et al.* (2018) 'Identification of Candida albicans regulatory genes governing mucosal infection', *Cellular microbiology*, 20(8), p. e12841.

Merry, C. R., Niland, C. and Khalil, A. M. (2015) 'Diverse Functions and Mechanisms of Mammalian Long Noncoding RNAs', *Methods in Molecular Biology*, pp. 1–14. doi: 10.1007/978-1-4939-1369-5_1.

Metpally, R. P. R. *et al.* (2013) 'Comparison of Analysis Tools for miRNA High Throughput Sequencing Using Nerve Crush as a Model', *Frontiers in genetics*, 4, p. 20.

Metzger, B. P. H., Wittkopp, P. J. and Coolon, J. D. (2017) 'Evolutionary Dynamics of Regulatory Changes Underlying Gene Expression Divergence among Saccharomyces Species', *Genome biology and evolution*, 9(4), pp. 843–854.

Mills, E. L., Kelly, B. and O’Neill, L. A. J. (2017) ‘Mitochondria are the powerhouses of immunity’, *Nature Immunology*, pp. 488–498. doi: 10.1038/ni.3704.

Mitrovich, Q. M. *et al.* (2007) ‘Computational and experimental approaches double the number of known introns in the pathogenic yeast Candida albicans’, *Genome research*, 17(4), pp. 492–502.

Mitsuhashi, S. *et al.* (2017) ‘A portable system for rapid bacterial composition analysis using a nanopore-based sequencer and laptop computer’, *Scientific reports*, 7(1), p. 5657.

Mixão, V. *et al.* (2019) ‘Whole-Genome Sequencing of the Opportunistic Yeast Pathogen Candida inconspicua Uncovers Its Hybrid Origin’, *Frontiers in Genetics*. doi: 10.3389/fgene.2019.00383.

Mixão, V. and Gabaldón, T. (2018) ‘Hybridization and emergence of virulence in opportunistic human yeast pathogens’, *Yeast* , 35(1), pp. 5–20.

Mixão, V. and Gabaldón, T. (2020) ‘Genomic evidence for a hybrid origin of the yeast opportunistic pathogen Candida albicans’, *BMC biology*, 18(1), p. 48.

Mohanty, A., Tiwari-Pandey, R. and Pandey, N. R. (2019) ‘Mitochondria: the indispensable players in innate immunity and guardians of the inflammatory response’, *Journal of cell communication and signaling*, 13(3), pp. 303–318.

Möller, M. *et al.* (2018) ‘Extraordinary Genome Instability and Widespread Chromosome Rearrangements During Vegetative Growth’, *Genetics*, 210(2), pp. 517–529.

Möller, M. and Stukenbrock, E. H. (2017) ‘Evolution and genome architecture in fungal plant pathogens’, *Nature reviews. Microbiology*, 15(12), pp. 756–771.

Monerawela, C. and Bond, U. (2018) ‘The hybrid genomes of Saccharomyces pastorianus: A current perspective’, *Yeast* , 35(1), pp. 39–50.

Morales, L. and Dujon, B. (2012) ‘Evolutionary role of interspecies hybridization and genetic exchanges in yeasts’, *Microbiology and molecular biology reviews: MMBR*, 76(4), pp. 721–739.

Moreno-Ruiz, E. *et al.* (2009) ‘Candida albicans internalization by host cells is mediated by a clathrin-dependent mechanism’, *Cellular microbiology*, 11(8), pp. 1179–1189.

Morrissy, A. S. *et al.* (2009) ‘Next-generation tag sequencing for cancer gene expression profiling’, *Genome research*, 19(10), pp. 1825–1835.

Mosser, D. M. and Edwards, J. P. (2008) ‘Exploring the full spectrum of macrophage activation’, *Nature reviews. Immunology*, 8(12), pp. 958–969.

Moyes, D. L. *et al.* (2010) ‘A biphasic innate immune MAPK response discriminates between the yeast and hyphal forms of Candida albicans in epithelial cells’, *Cell host & microbe*, 8(3), pp. 225–235.

Moyes, D. L. *et al.* (2011) ‘Candida albicans Yeast and Hyphae are Discriminated by MAPK Signaling in Vaginal Epithelial Cells’, *PLoS ONE*, p. e26580. doi: 10.1371/journal.pone.0026580.

Moyes, D. L. *et al.* (2016) ‘Candidalysin is a fungal peptide toxin critical for mucosal infection’, *Nature*, 532(7597), pp. 64–68.

Moyes, D. L. and Naglik, J. R. (2011) ‘Mucosal immunity and Candida albicans infection’, *Clinical & developmental immunology*, 2011, p. 346307.

Munakata, K. *et al.* (2008) ‘Importance of the interferon-alpha system in murine large intestine indicated by microarray analysis of commensal bacteria-induced immunological changes’, *BMC genomics*, 9, p. 192.

Muzzey, D., Sherlock, G. and Weissman, J. S. (2014) ‘Extensive and coordinated control of allele-specific expression by both transcription and translation in Candida albicans’, *Genome research*, 24(6), pp. 963–973.

Nagalakshmi, U. *et al.* (2008) ‘The transcriptional landscape of the yeast genome defined by RNA sequencing’, *Science*, 320(5881), pp. 1344–1349.

Naglik, J. R. *et al.* (2011) ‘Candida albicans interactions with epithelial cells and mucosal immunity’, *Microbes and infection / Institut Pasteur*, 13(12-13), pp. 963–976.

Naglik, J. R., Challacombe, S. J. and Hube, B. (2003) ‘Candida albicans secreted aspartyl proteinases in virulence and pathogenesis’, *Microbiology and molecular biology reviews: MMBR*, 67(3), pp. 400–28, table of contents.

Naglik, J. R. and Moyes, D. (2011) ‘Epithelial Cell Innate Response to Candida albicans’, *Advances in Dental Research*, pp. 50–55. doi: 10.1177/0022034511399285.

Naranjo-Ortiz, M. A. and Gabaldón, T. (2019) ‘Fungal evolution: diversity, taxonomy and phylogeny of the Fungi’, *Biological Reviews*, pp. 2101–2137. doi: 10.1111/brv.12550.

Németh, T. *et al.* (2013) ‘Characterization of Virulence Properties in the C. parapsilosis Sensu Lato Species’, *PLoS ONE*, p. e68704. doi: 10.1371/journal.pone.0068704.

Neofytos, D. *et al.* (2013) ‘Epidemiology, outcomes, and risk factors of invasive fungal infections in adult patients with acute myelogenous leukemia after induction chemotherapy’, *Diagnostic microbiology and infectious disease*, 75(2), pp. 144–149.

Netea, M. G. *et al.* (2008) ‘An integrated model of the recognition of Candida albicans by the innate immune system’, *Nature reviews. Microbiology*, 6(1), pp. 67–78.

Niederer, R. O., Hass, E. P. and Zappulla, D. C. (2017) ‘Long Noncoding RNAs in the Yeast S. cerevisiae’, *Advances in experimental medicine and biology*, 1008, pp. 119–132.

Niemiec, M. J. *et al.* (2017) ‘Dual transcriptome of the immediate neutrophil and Candida albicans interplay’, *BMC genomics*, 18(1), p. 696.

Ning, G. *et al.* (2017) ‘Hybrid sequencing and map finding (HySeMaFi): optional strategies for extensively deciphering gene splicing and expression in organisms without reference genome’, *Scientific reports*, 7, p. 43793.

Ning, X. *et al.* (2019) ‘Apoptotic Caspases Suppress Type I Interferon Production via the Cleavage of cGAS, MAVS, and IRF3’, *Molecular cell*, 74(1), pp. 19–31.e7.

Nobile, C. J. *et al.* (2006) ‘Function of Candida albicans adhesin Hwp1 in biofilm formation’, *Eukaryotic cell*, 5(10), pp. 1604–1610.

Noble, S. M., Gianetti, B. A. and Witchley, J. N. (2017) ‘Candida albicans cell-type switching and functional plasticity in the mammalian host’, *Nature Reviews Microbiology*, pp. 96–108. doi:

10.1038/nrmicro.2016.157.

Nomaguchi, T. *et al.* (2018) ‘Homoeolog expression bias in allopolyploid oleaginous marine diatom Fistulifera solaris’, *BMC genomics*, 19(1), p. 330.

Novačić, A. *et al.* (2020) ‘Non-coding RNAs as cell wall regulators in Saccharomyces cerevisiae’, *Critical Reviews in Microbiology*, pp. 15–25. doi: 10.1080/1040841x.2020.1715340.

Nuss, A. M. *et al.* (2017) ‘Tissue dual RNA-seq allows fast discovery of infection-specific functions and riboregulators shaping host-pathogen transcriptomes’, *Proceedings of the National Academy of Sciences of the United States of America*, 114(5), pp. E791–E800.

O’Brien, H. E. *et al.* (2005) ‘Fungal community analysis by large-scale sequencing of environmental samples’, *Applied and environmental microbiology*, 71(9), pp. 5544–5550.

O’Keeffe, G. *et al.* (2014) ‘RNA-seq reveals the pan-transcriptomic impact of attenuating the gliotoxin self-protection mechanism in Aspergillus fumigatus’, *BMC genomics*, 15, p. 894.

O’Leary, N. A. *et al.* (2016) ‘Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation’, *Nucleic Acids Research*, pp. D733–D745. doi: 10.1093/nar/gkv1189.

O’Meara, T. R. *et al.* (2010) ‘Interaction of Cryptococcus neoformans Rim101 and protein kinase A regulates capsule’, *PLoS pathogens*, 6(2), p. e1000776.

O’Meara, T. R. *et al.* (2013) ‘Cryptococcus neoformans Rim101 is associated with cell wall remodeling and evasion of the host immune responses’, *mBio*, 4(1). doi: 10.1128/mBio.00522-12.

O’Neil, D., Glowatz, H. and Schlumpberger, M. (2013) ‘Ribosomal RNA depletion for efficient use of RNA-seq capacity’, *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, Chapter 4, p. Unit 4.19.

Oren, I. and Paul, M. (2014) ‘Up to date epidemiology, diagnosis and management of invasive fungal infections’, *Clinical microbiology and infection: the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, 20 Suppl 6, pp. 1–4.

Orsi, W., Biddle, J. F. and Edgcomb, V. (2013) ‘Deep sequencing of subseafloor eukaryotic rRNA reveals active Fungi across marine subsurface provinces’, *PloS one*, 8(2), p. e56335.

Otto, C., Stadler, P. F. and Hoffmann, S. (2014) ‘Lacking alignments? The next-generation sequencing mapper segemehl revisited’, *Bioinformatics* , 30(13), pp. 1837–1843.

Ouyang, J., Hu, J. and Chen, J.-L. (2016) ‘lncRNAs regulate the innate immune response to viral infection’, *Wiley interdisciplinary reviews. RNA*, 7(1), pp. 129–143.

Pammi, M. *et al.* (2013) ‘Candida parapsilosis is a significant neonatal pathogen: a systematic review and meta-analysis’, *The Pediatric infectious disease journal*, 32(5), pp. e206–16.

Papon, N. *et al.* (2013) ‘Emerging and emerged pathogenic Candida species: beyond the Candida albicans paradigm’, *PLoS pathogens*, 9(9), p. e1003550.

Paralkar, V. R. *et al.* (2014) ‘Lineage and species-specific long noncoding RNAs during erythro-megakaryocytic development’, *Blood*, 123(12), pp. 1927–1937.

Park, B. J. *et al.* (2009) ‘Estimation of the current global burden of cryptococcal meningitis among persons living with HIV/AIDS’, *AIDS* , 23(4), pp. 525–530.

Parkhomchuk, D. *et al.* (2009) ‘Transcriptome analysis by strand-specific sequencing of complementary DNA’, *Nucleic acids research*, 37(18), p. e123.

Patel, R. K. and Jain, M. (2012) ‘NGS QC Toolkit: a toolkit for quality control of next generation sequencing data’, *PloS one*, 7(2), p. e30619.

Paterson, R. R. M. and Lima, N. (2017) ‘Filamentous Fungal Human Pathogens from Food Emphasising Aspergillus, Fusarium and Mucor’, *Microorganisms*, 5(3). doi: 10.3390/microorganisms5030044.

Patro, R. *et al.* (2017) ‘Salmon provides fast and bias-aware quantification of transcript expression’, *Nature methods*, 14(4), pp. 417–419.

Paun, O. *et al.* (2007) ‘Genetic and epigenetic alterations after hybridization and genome doubling’, *Taxon*, 56(3), pp. 649–656.

Pegueroles, C., Mixão, V., *et al.* (2019) ‘HaploTypo: a variant-calling pipeline for phased genomes’, *Bioinformatics* . doi: 10.1093/bioinformatics/btz933.

Pegueroles, C., Iraola-Guzmán, S., *et al.* (2019) ‘Transcriptomic analyses reveal groups of co-expressed, syntenic lncRNAs in four species of the genus Caenorhabditis’, *RNA biology*, 16(3), pp. 320–329.

Pegueroles, C. and Gabaldón, T. (2016) ‘Secondary structure impacts patterns of selection in human lncRNAs’, *BMC biology*, 14, p. 60.

Pekmezovic, M. *et al.* (2019) ‘Host-Pathogen Interactions during Female Genital Tract Infections’, *Trends in microbiology*, 27(12), pp. 982–996.

Pelechano, V. and Pérez-Ortín, J. E. (2008) ‘The transcriptional inhibitor thiolutin blocks mRNA degradation in yeast’, *Yeast* , 25(2), pp. 85–92.

Pel, H. J. *et al.* (2007) ‘Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88’, *Nature biotechnology*, 25(2), pp. 221–231.

Pérez-Torrado, R. *et al.* (2015) ‘Molecular and enological characterization of a natural Saccharomyces uvarum and Saccharomyces cerevisiae hybrid’, *International journal of food microbiology*, 204, pp. 101–110.

Perfect, J. R. *et al.* (2001) ‘The impact of culture isolation of Aspergillus species: a hospital-based survey of aspergillosis’, *Clinical infectious diseases: an official publication of the Infectious Diseases Society of America*, 33(11), pp. 1824–1833.

Pertea, M. *et al.* (2015) ‘StringTie enables improved reconstruction of a transcriptome from RNA-seq reads’, *Nature biotechnology*, 33(3), pp. 290–295.

Pervolaraki, K. *et al.* (2018) ‘Differential induction of interferon stimulated genes between type I and type III interferons is independent of interferon receptor abundance’, *PLoS pathogens*, 14(11), p. e1007420.

Petersen, T. N. *et al.* (2017) ‘MGmapper: Reference based mapping and taxonomy annotation of metagenomics sequence reads’, *PloS one*, 12(5), p. e0176469.

Pfaller, M. A. *et al.* (2009) ‘Variation in susceptibility of bloodstream isolates of Candida glabrata to fluconazole according to patient age and geographic location in the United States in 2001 to 2007’, *Journal of* *clinical microbiology*, 47(10), pp. 3185–3190.

Pfaller, M. A. and Diekema, D. J. (2007) ‘Epidemiology of invasive candidiasis: a persistent public health problem’, *Clinical microbiology reviews*, 20(1), pp. 133–163.

Pfaller, M. A. and Diekema, D. J. (2010) ‘Epidemiology of Invasive Mycoses in North America’, *Critical Reviews in Microbiology*, pp. 1–53. doi: 10.3109/10408410903241444.

Pirofski, L.-A. and Casadevall, A. (2008) ‘The Damage-Response Framework of Microbial Pathogenesis and Infectious Diseases’, *Advances in Experimental Medicine and Biology*, pp. 135–146. doi: 10.1007/978-0-387-09550-9_11.

Plataki, M. *et al.* (2019) ‘Mitochondrial Dysfunction in Aged Macrophages and Lung during Primary Streptococcus pneumoniae Infection is Improved with Pirfenidone’, *Scientific Reports*. doi: 10.1038/s41598-018-37438-1.

Plissonneau, C., Stürchler, A. and Croll, D. (2016) ‘The Evolution of Orphan Regions in Genomes of a Fungal Pathogen of Wheat’, *mBio*, 7(5). doi: 10.1128/mBio.01231-16.

Pruitt, K. D., Tatusova, T. and Maglott, D. R. (2007) ‘NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins’, *Nucleic acids research*, 35(Database issue), pp. D61–5.

Pryszcz, L. P. *et al.* (2014) ‘Genome comparison of Candida orthopsilosis clinical strains reveals the existence of hybrids between two distinct subspecies’, *Genome biology and evolution*, 6(5), pp. 1069–1078.

Pryszcz, L. P. *et al.* (2015) ‘The Genomic Aftermath of Hybridization in the Opportunistic Pathogen Candida metapsilosis’, *PLoS genetics*, 11(10), p. e1005626.

Quick, J. *et al.* (2015) ‘Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella’, *Genome biology*, 16, p. 114.

Quince, C. *et al.* (2017) ‘Corrigendum: Shotgun metagenomics, from sampling to analysis’, *Nature biotechnology*, 35(12), p. 1211.

Quinlan, A. R. (2014) ‘BEDTools: The Swiss-Army Tool for Genome

Feature Analysis', *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, 47, pp. 11.12.1–34.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics* , 26(6), pp. 841–842.

Ramond, E. *et al.* (2019) 'Pivotal Role of Mitochondria in Macrophage Response to Bacterial Pathogens', *Frontiers in Immunology*. doi: 10.3389/fimmu.2019.02461.

Rando, O. J. and Chang, H. Y. (2009) 'Genome-wide views of chromatin structure', *Annual review of biochemistry*, 78, pp. 245–271.

Rapaport, F. *et al.* (2013) 'Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data', *Genome biology*, 14(9), p. R95.

Rappolee, D. A. *et al.* (1988) 'Wound macrophages express TGF-alpha and other growth factors in vivo: analysis by mRNA phenotyping', *Science*, 241(4866), pp. 708–712.

Rasheed, M., Battu, A. and Kaur, R. (2018) 'Aspartyl proteases in are required for suppression of the host innate immune response', *The Journal of biological chemistry*, 293(17), pp. 6410–6433.

Reuter, S. *et al.* (2013) 'Rapid bacterial whole-genome sequencing to enhance diagnostic and public health microbiology', *JAMA internal medicine*, 173(15), pp. 1397–1404.

Rhoads, A. and Au, K. F. (2015) 'PacBio Sequencing and Its Applications', *Genomics, proteomics & bioinformatics*, 13(5), pp. 278–289.

Rhodes, J. *et al.* (2017) 'Tracing Genetic Exchange and Biogeography of var. at the Global Population Level', *Genetics*, 207(1), pp. 327–346.

Riccombeni, A. *et al.* (2012) 'Sequence and analysis of the genome of the pathogenic yeast Candida orthopsilosis', *PloS one*, 7(4), p. e35750.

Richardson, J. P. *et al.* (2018) 'Candidalysin Drives Epithelial Signaling, Neutrophil Recruitment, and Immunopathology at the Vaginal Mucosa', *Infection and immunity*, 86(2). doi: 10.1128/IAI.00645-17.

Richardson, J. P., Ho, J. and Naglik, J. R. (2018) 'Candida-Epithelial

Interactions’, *Journal of fungi (Basel, Switzerland),* 4(1). doi: 10.3390/jof4010022.

Riedelberger, M. *et al.* (2020) ‘Type I Interferon Response Dysregulates Host Iron Homeostasis and Enhances Candida glabrata Infection’, *Cell host & microbe,* 27(3), pp. 454–466.e8.

Rieseberg, L. H. (1997) ‘HYBRID ORIGINS OF PLANT SPECIES’, *Annual Review of Ecology and Systematics*, pp. 359–389. doi: 10.1146/annurev.ecolsys.28.1.359.

Riley, J. S. and Tait, S. W. (2020) ‘Mitochondrial DNA in inflammation and immunity’, *EMBO reports,* 21(4), p. e49799.

Risso, D. *et al.* (2014) ‘Normalization of RNA-seq data using factor analysis of control genes or samples’, *Nature Biotechnology,* pp. 896–902. doi: 10.1038/nbt.2931.

Ritchie, M. E. *et al.* (2015) ‘limma powers differential expression analyses for RNA-sequencing and microarray studies’, *Nucleic acids research,* 43(7), p. e47.

Rizzetto, L. *et al.* (2013) ‘Strain dependent variation of immune responses to A. fumigatus: definition of pathogenic species’, *PloS one,* 8(2), p. e56651.

Robert, V. A. and Casadevall, A. (2009) ‘Vertebrate endothermy restricts most fungi as potential pathogens’, *The Journal of infectious diseases,* 200(10), pp. 1623–1626.

Robinson, J. T. *et al.* (2011) ‘Integrative genomics viewer’, *Nature Biotechnology,* pp. 24–26. doi: 10.1038/nbt.1754.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010) ‘edgeR: a Bioconductor package for differential expression analysis of digital gene expression data’, *Bioinformatics* , 26(1), pp. 139–140.

Rodrigues, C. F. *et al.* (2017) ‘Candida glabrata Biofilms: How Far Have We Come?’, *Journal of fungi (Basel, Switzerland),* 3(1). doi: 10.3390/jof3010011.

Rodríguez, A. *et al.* (2019) ‘Nucleic acids enrichment of fungal pathogens to study host-pathogen interactions’, *Scientific reports,* 9(1), p. 18037.

Rodriguez, R. J. and Redman, R. S. (1997) ‘Fungal Life-Styles and

Ecosystem Dynamics: Biological Aspects of Plant Pathogens, Plant Endophytes and Saprophytes', *Advances in Botanical Research*, pp. 169–193. doi: 10.1016/s0065-2296(08)60073-7.

Romani, L. (2011) 'Immunity to fungal infections', *Nature reviews. Immunology*, 11(4), pp. 275–288.

Ropars, J. *et al.* (2018) 'Gene flow contributes to diversification of the major fungal pathogen Candida albicans', *Nature communications*, 9(1), p. 2253.

Rosati, D. *et al.* (2020) 'Recurrent Vulvovaginal Candidiasis: An Immunological Perspective', *Microorganisms*, 8(2). doi: 10.3390/microorganisms8020144.

Rosenbach, A. *et al.* (2010) 'Adaptations of Candida albicans for growth in the mammalian intestinal tract', *Eukaryotic cell*, 9(7), pp. 1075–1086.

Rosenberg, A. B. *et al.* (2018) 'Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding', *Science*, 360(6385), pp. 176–182.

Rosenthal, K. *et al.* (2017) 'Beyond the bulk: disclosing the life of single microbial cells', *FEMS microbiology reviews*, 41(6), pp. 751–780.

Ross-Innes, C. S. *et al.* (2012) 'Differential oestrogen receptor binding is associated with clinical outcome in breast cancer', *Nature*, 481(7381), pp. 389–393.

Ross, M. G. *et al.* (2013) 'Characterizing and measuring bias in sequence data', *Genome biology*, 14(5), p. R51.

Salazar, F. and Brown, G. D. (2018) 'Antifungal Innate Immunity: A Perspective from the Last 10 Years', *Journal of innate immunity*, 10(5-6), pp. 373–397.

Saliba, A.-E. *et al.* (2016) 'Single-cell RNA-seq ties macrophage polarization to growth rate of intracellular Salmonella', *Nature microbiology*, 2, p. 16206.

Salinas, F. *et al.* (2016) 'Natural variation in non-coding regions underlying phenotypic diversity in budding yeast', *Scientific reports*, 6, p. 21849.

Salvadó, Z. *et al.* (2011) 'Temperature adaptation markedly determines evolution within the genus Saccharomyces', *Applied and environmental microbiology*, 77(7), pp. 2292–2302.

Samson, R. A. *et al.* (2014) 'Phylogeny, identification and nomenclature of the genus Aspergillus', *Studies in mycology*, 78, pp. 141–173.

Sanger, F., Nicklen, S. and Coulson, A. R. (1992) 'DNA sequencing with chain-terminating inhibitors. 1977', *Biotechnology* , 24, pp. 104–108.

Sanglard, D. (2016) 'Emerging Threats in Antifungal-Resistant Fungal Pathogens', *Frontiers of medicine*, 3, p. 11.

Santos, M. A. S. *et al.* (2011) 'The genetic code of the fungal CTG clade', *Comptes rendus biologies*, 334(8-9), pp. 607–611.

Sarma, S. and Upadhyay, S. (2017) 'Current perspective on emergence, diagnosis and drug resistance in Candida auris', *Infection and Drug Resistance*, pp. 155–165. doi: 10.2147/idr.s116229.

Sarropoulos, I. *et al.* (2019) 'Developmental dynamics of lncRNAs across mammalian organs and species', *Nature*, 571(7766), pp. 510–514.

Satoh, K. *et al.* (2009) 'Candida auris sp. nov., a novel ascomycetous yeast isolated from the external ear canal of an inpatient in a Japanese hospital', *Microbiology and immunology*, 53(1), pp. 41–44.

Sato, M. *et al.* (1998) 'Positive feedback regulation of type I IFN genes by the IFN-inducible transcription factor IRF-7', *FEBS letters*, 441(1), pp. 106–110.

Saus, E. *et al.* (2018) 'nextPARS: parallel probing of RNA structures in Illumina', *RNA* , 24(4), pp. 609–619.

Scannell, D. R. *et al.* (2011) 'The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus', *G3: Genes|Genomes|Genetics*, 1(1), pp. 11–25.

Schaefke, B. *et al.* (2013) 'Inheritance of gene expression level and selective constraints on trans- and cis-regulatory changes in yeast', *Molecular biology and evolution*, 30(9), pp. 2121–2133.

Schaller, M. and Weindl, G. (2009) 'Models of oral and vaginal candidiasis based on in vitro reconstituted human epithelia for the study of host-pathogen interactions', *Methods in molecular biology* , 470, pp. 327–

345.

Scheele, B. C. *et al.* (2019) ‘Amphibian fungal panzootic causes catastrophic and ongoing loss of biodiversity’, *Science*, 363(6434), pp. 1459–1463.

Schena, M. *et al.* (1995) ‘Quantitative monitoring of gene expression patterns with a complementary DNA microarray’, *Science*, 270(5235), pp. 467–470.

Schep, A. N. *et al.* (2015) ‘Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions’, *Genome research*, 25(11), pp. 1757–1770.

Schmidt, K. *et al.* (2017) ‘Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing’, *The Journal of antimicrobial chemotherapy*, 72(1), pp. 104–114.

Schneider, W. M., Chevillotte, M. D. and Rice, C. M. (2014) ‘Interferon-stimulated genes: a complex web of host defenses’, *Annual review of immunology*, 32, pp. 513–545.

Schor, I. E. *et al.* (2018) ‘Non-coding RNA Expression, Function, and Variation during Drosophila Embryogenesis’, *Current biology: CB*, 28(22), pp. 3547–3561.e9.

Schraiber, J. G. *et al.* (2013) ‘Inferring evolutionary histories of pathway regulation from transcriptional profiling data’, *PLoS computational biology*, 9(10), p. e1003255.

Schröder, M. S., de San Vicente, K. M., *et al.* (2016) ‘Multiple Origins of the Pathogenic Yeast Candida orthopsilosis by Separate Hybridizations between Two Parental Species’, *PLOS Genetics*, p. e1006404. doi: 10.1371/journal.pgen.1006404.

Schröder, M. S., Martinez de San Vicente, K., *et al.* (2016) ‘Multiple Origins of the Pathogenic Yeast Candida orthopsilosis by Separate Hybridizations between Two Parental Species’, *PLoS genetics*, 12(11), p. e1006404.

Schulze, S. *et al.* (2015) ‘Computational prediction of molecular pathogen-host interactions based on dual transcriptome data’, *Frontiers in microbiology*, 6, p. 65.

Schulze, S. *et al.* (2016) ‘How to Predict Molecular Interactions between Species?’, *Frontiers in microbiology*, 7, p. 442.

Schurch, N. J. *et al.* (2016) ‘How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?’, *RNA* , 22(6), pp. 839–851.

Schwenk, K., Brede, N. and Streit, B. (2008) ‘Introduction. Extent, processes and evolutionary impact of interspecific hybridization in animals’, *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 363(1505), pp. 2805–2811.

Segal, A. W. (2005) ‘HOW NEUTROPHILS KILL MICROBES’, *Annual Review of Immunology*, pp. 197–223. doi: 10.1146/annurev.immunol.23.021704.115653.

Segal, E. and Frenkel, M. (2018) ‘Experimental in Vivo Models of Candidiasis’, *Journal of fungi (Basel, Switzerland)*, 4(1). doi: 10.3390/jof4010021.

Selmecki, A., Bergmann, S. and Berman, J. (2005) ‘Comparative genome hybridization reveals widespread aneuploidy in Candida albicans laboratory strains’, *Molecular microbiology*, 55(5), pp. 1553–1565.

Selmecki, A. M. *et al.* (2009) ‘Acquisition of aneuploidy provides increased fitness during the evolution of antifungal drug resistance’, *PLoS genetics*, 5(10), p. e1000705.

SEQC/MAQC-III Consortium (2014) ‘A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium’, *Nature biotechnology*, 32(9), pp. 903–914.

Serin, E. A. R. *et al.* (2016) ‘Learning from Co-expression Networks: Possibilities and Challenges’, *Frontiers in plant science*, 7, p. 444.

Session, A. M. *et al.* (2016) ‘Genome evolution in the allotetraploid frog Xenopus laevis’, *Nature*, 538(7625), pp. 336–343.

Seth, R. B. *et al.* (2005) ‘Identification and Characterization of MAVS, a Mitochondrial Antiviral Signaling Protein that Activates NF-κB and IRF3’, *Cell*, pp. 669–682. doi: 10.1016/j.cell.2005.08.012.

Seyednasrollah, F., Laiho, A. and Elo, L. L. (2015) ‘Comparison of software packages for detecting differential expression in RNA-seq studies’, *Briefings in bioinformatics*, 16(1), pp. 59–70.

Shankar, J. *et al.* (2018) ‘RNA-Seq Profile Reveals Th-1 and Th-17-Type of Immune Responses in Mice Infected Systemically with Aspergillus fumigatus’, *Mycopathologia*, 183(4), pp. 645–658.

Sharon, D. *et al.* (2013) ‘A single-molecule long-read survey of the human transcriptome’, *Nature biotechnology*, 31(11), pp. 1009–1014.

Shaw, W. H. *et al.* (2016) ‘Identification of HIV Mutation as Diagnostic Biomarker through Next Generation Sequencing’, *Journal of clinical and diagnostic research: JCDR*, 10(7), pp. DC04–8.

Shen, X.-X. *et al.* (2018) ‘Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum’, *Cell*, 175(6), pp. 1533–1545.e20.

Sherry, N. L. *et al.* (2013) ‘Outbreak investigation using high-throughput genome sequencing within a diagnostic microbiology laboratory’, *Journal of clinical microbiology*, 51(5), pp. 1396–1401.

Shimada, K. *et al.* (2012) ‘Oxidized mitochondrial DNA activates the NLRP3 inflammasome during apoptosis’, *Immunity*, 36(3), pp. 401–414.

Short, D. P. G., O’Donnell, K. and 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, p. 91.

Skrzypek, M. S. *et al.* (2017) ‘The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data’, *Nucleic acids research*, 45(D1), pp. D592–D596.

Smeekens, S. P. *et al.* (2013) ‘Functional genomics identifies type I interferon pathway as central for host defense against Candida albicans’, *Nature Communications*. doi: 10.1038/ncomms2343.

Smeekens, S. P., van de Veerdonk, F. L. and Netea, M. G. (2016) ‘An Omics Perspective on Candida Infections: Toward Next-Generation Diagnosis and Therapy’, *Frontiers in microbiology*, 7, p. 154.

Smukowski Heil, C. S. *et al.* (2017) ‘Loss of Heterozygosity Drives Adaptation in Hybrid Yeast’, *Molecular biology and evolution*, 34(7), pp. 1596–1612.

Sobel, J. D. (2007) ‘Vulvovaginal candidosis’, *The Lancet*, 369(9577), pp. 1961–1971.

Soneson, C. and Delorenzi, M. (2013) ‘A comparison of methods for differential expression analysis of RNA-seq data’, *BMC bioinformatics*, 14, p. 91.

Soneson, C., Love, M. I. and Robinson, M. D. (2015) ‘Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences’, *F1000Research*, 4, p. 1521.

Spitale, R. C. *et al.* (2015) ‘Structural imprints in vivo decode RNA regulatory mechanisms’, *Nature*, 519(7544), pp. 486–490.

Stajich, J. E. *et al.* (2012) ‘FungiDB: an integrated functional genomics database for fungi’, *Nucleic acids research*, 40(Database issue), pp. D675–81.

Stavru, F. *et al.* (2011) ‘Listeria monocytogenes transiently alters mitochondrial dynamics during infection’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(9), pp. 3612–3617.

Steenwyk, J. L. and Rokas, A. (2018) ‘Copy Number Variation in Fungi and Its Implications for Wine Yeast Genetic Diversity and Adaptation’, *Frontiers in microbiology*, 9, p. 288.

Stegle, O., Teichmann, S. A. and Marioni, J. C. (2015) ‘Computational and analytical challenges in single-cell transcriptomics’, *Nature reviews. Genetics*, 16(3), pp. 133–145.

Stoesser, N. *et al.* (2013) ‘Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data’, *The Journal of antimicrobial chemotherapy*, 68(10), pp. 2234–2244.

‘Stop neglecting fungi’ (2017) *Nature microbiology*, 2, p. 17120.

Stukenbrock, E. H. *et al.* (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*, pp. 10954–10959. doi: 10.1073/pnas.1201403109.

Stukenbrock, E. H. (2016) ‘The Role of Hybridization in the Evolution and Emergence of New Fungal Plant Pathogens’, *Phytopathology*, pp.

104–112. doi: 10.1094/phyto-08-15-0184-rvw.

Subramanian, A. *et al.* (2005) ‘Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles’, *Proceedings of the National Academy of Sciences of the United States of America*, 102(43), pp. 15545–15550.

Sudbery, P., Gow, N. and Berman, J. (2004) ‘The distinct morphogenic states of Candida albicans’, *Trends in microbiology*, 12(7), pp. 317–324.

Sullivan, D. J. and Moran, G. P. (2014) *Human Pathogenic Fungi: Molecular Biology and Pathogenic Mechanisms*.

Sundstrom, P. (1999) ‘Adhesins in Candida albicans’, *Current opinion in microbiology*, 2(4), pp. 353–357.

Sun, W.-H. *et al.* (2019) ‘Genome-wide analysis of long non-coding RNAs in Pichia pastoris during stress by RNA sequencing’, *Genomics*, 111(3), pp. 398–406.

Sutcliffe, J. G. *et al.* (1982) ‘Common 82-nucleotide sequence unique to brain RNA’, *Proceedings of the National Academy of Sciences of the United States of America*, 79(16), pp. 4942–4946.

Swidergall, M. *et al.* (2018) ‘EphA2 is an epithelial cell pattern recognition receptor for fungal β-glucans’, *Nature microbiology*, 3(1), pp. 53–61.

Syn, G. *et al.* (2017) ‘Toxoplasma gondii Infection Is Associated with Mitochondrial Dysfunction in-Vitro’, *Frontiers in Cellular and Infection Microbiology*. doi: 10.3389/fcimb.2017.00512.

Taei, M., Chadeganipour, M. and Mohammadi, R. (2019) ‘An alarming rise of non-albicans Candida species and uncommon yeasts in the clinical samples; a combination of various molecular techniques for identification of etiologic agents’, *BMC research notes*, 12(1), p. 779.

Tarazona, S. *et al.* (2011) ‘Differential expression in RNA-seq: a matter of depth’, *Genome research*, 21(12), pp. 2213–2223.

Tavanti, A. *et al.* (2005) ‘Candida orthopsilosis and Candida metapsilosis spp. nov. to replace Candida parapsilosis groups II and III’, *Journal of clinical microbiology*, 43(1), pp. 284–292.

Teixeira, M. C. *et al.* (2018) ‘YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in Saccharomyces cerevisiae’, *Nucleic acids research*, 46(D1), pp. D348–D353.

Thänert, R. *et al.* (2017) ‘Host-inherent variability influences the transcriptional response of Staphylococcus aureus during in vivo infection’, *Nature communications*, 8, p. 14268.

The Gene Ontology Consortium (2017) ‘Expansion of the Gene Ontology knowledgebase and resources’, *Nucleic acids research*, 45(D1), pp. D331–D338.

Thewes, S. *et al.* (2007) ‘In vivo and ex vivo comparative transcriptional profiling of invasive and non-invasive Candida albicans isolates identifies genes associated with tissue invasion’, *Molecular microbiology*, 63(6), pp. 1606–1628.

Thompson, D. A. and Cubillos, F. A. (2017) ‘Natural gene expression variation studies in yeast’, *Yeast* , 34(1), pp. 3–17.

Thompson, D. A. and Regev, A. (2009) ‘Fungal regulatory evolution: cis and trans in the balance’, *FEBS letters*, 583(24), pp. 3959–3965.

Thuer, E. (2017) *Deep Sequencing Approaches to Investigate the Dynamics and Evolution of Interaction Networks of Candida Pathogens and the Human Host.*

Tierney, L. *et al.* (2012) ‘An Interspecies Regulatory Network Inferred from Simultaneous RNA-seq of Candida albicans Invading Innate Immune Cells’, *Frontiers in microbiology*, 3, p. 85.

Till, P., Mach, R. L. and Mach-Aigner, A. R. (2018) ‘A current view on long noncoding RNAs in yeast and filamentous fungi’, *Applied microbiology and biotechnology*, 102(17), pp. 7319–7331.

Tipper, D. J. (1973) ‘Inhibition of yeast ribonucleic acid polymerases by thiolutin’, *Journal of bacteriology*, 116(1), pp. 245–256.

Tirosh, I. *et al.* (2009) ‘A yeast hybrid provides insight into the evolution of gene expression regulation’, *Science*, 324(5927), pp. 659–662.

Tirosh, I., Sigal, N. and Barkai, N. (2010) ‘Divergence of nucleosome positioning between two closely related yeast species: genetic basis and functional consequences’, *Molecular systems biology*, 6, p. 365.

Tisseur, M., Kwapisz, M. and Morillon, A. (2011) ‘Pervasive transcription

– Lessons from yeast', *Biochimie*, pp. 1889–1896. doi: 10.1016/j.biochi.2011.07.001.

Todd, R. T., Forche, A. and Selmecki, A. (2017) 'Ploidy Variation in Fungi: Polyploidy, Aneuploidy, and Genome Evolution', *Microbiology Spectrum*. doi: 10.1128/microbiolspec.funk-0051-2016.

Tóth, R. *et al.* (2018) 'Investigation of Candida parapsilosis virulence regulatory factors during host-pathogen interaction', *Scientific reports*, 8(1), p. 1346.

Tóth, R. *et al.* (2019) 'Candida parapsilosis: from Genes to the Bedside', *Clinical microbiology reviews*, 32(2). doi: 10.1128/CMR.00111-18.

Trapnell, C. *et al.* (2010) 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation', *Nature biotechnology*, 28(5), pp. 511–515.

Trapnell, C. *et al.* (2012) 'Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks', *Nature protocols*, 7(3), pp. 562–578.

Turabelidze, G. *et al.* (2013) 'Precise dissection of an Escherichia coli O157:H7 outbreak by single nucleotide polymorphism analysis', *Journal of clinical microbiology*, 51(12), pp. 3950–3954.

Turner, S. A. and Butler, G. (2014) 'The Candida pathogenic species complex', *Cold Spring Harbor perspectives in medicine*, 4(9), p. a019778.

Uszczynska-Ratajczak, B. *et al.* (2018) 'Towards a complete map of the human long non-coding RNA transcriptome', *Nature reviews. Genetics*, 19(9), pp. 535–548.

Velculescu, V. E. *et al.* (1995) 'Serial analysis of gene expression', *Science*, 270(5235), pp. 484–487.

Verma, A., Gaffen, S. and Swidergall, M. (2017) 'Innate Immunity to Mucosal Candida Infections', *Journal of Fungi*, p. 60. doi: 10.3390/jof3040060.

'Virus interference. I. The interferon' (1957) *Proceedings of the Royal Society of London. Series B - Biological Sciences*, pp. 258–267. doi: 10.1098/rspb.1957.0048.

Wächtler, B. *et al.* (2011) 'From attachment to damage: defined genes of

Candida albicans mediate adhesion, invasion and damage during interaction with oral epithelial cells', *PloS one*, 6(2), p. e17046.

Wächtler, B. *et al.* (2012) 'Candida albicans-epithelial interactions: dissecting the roles of active penetration, induced endocytosis and host factors on the infection process', *PloS one*, 7(5), p. e36952.

Wächtler, B., Wilson, D. and Hube, B. (2011) 'Candida albicans adhesion to and invasion and damage of vaginal epithelial cells: stage-specific inhibition by clotrimazole and bifonazole', *Antimicrobial agents and chemotherapy*, 55(9), pp. 4436–4439.

Wain, J. and Mavrogiorgou, E. (2013) 'Next-generation sequencing in clinical microbiology', *Expert review of molecular diagnostics*, 13(3), pp. 225–227.

Wainwright, M. *et al.* (2003) 'Microorganisms cultured from stratospheric air samples obtained at 41 km', *FEMS microbiology letters*, 218(1), pp. 161–165.

Walker, L. A. *et al.* (2009) 'Genome-wide analysis of Candida albicans gene expression patterns during infection of the mammalian kidney', *Fungal genetics and biology: FG & B*, 46(2), pp. 210–219.

Wang, B. *et al.* (2018) 'A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing', *Genome research*, 28(6), pp. 921–932.

Wang, J. *et al.* (2015) 'RNA-seq based transcriptomic analysis of single bacterial cells', *Integrative biology: quantitative biosciences from nano to macro*, 7(11), pp. 1466–1476.

Wang, K. *et al.* (2015) 'Transcription factor ADS-4 regulates adaptive responses and resistance to antifungal azole stress', *Antimicrobial agents and chemotherapy*, 59(9), pp. 5396–5404.

Wang, P.-H. *et al.* (2018) 'A novel transcript isoform of STING that sequesters cGAMP and dominantly inhibits innate nucleic acid sensing', *Nucleic acids research*, 46(8), pp. 4054–4071.

Wang, Y. *et al.* (2019) 'XRN1-associated long non-coding RNAs may contribute to fungal virulence and sexual development in entomopathogenic fungus Cordyceps militaris', *Pest management science*, 75(12), pp. 3302–3311.

Wang, Z. *et al.* (2019) ‘Genome-Wide Identification and Functional Prediction of Long Non-coding RNAs Involved in the Heat Stress Response in’, *Frontiers in microbiology*, 10, p. 2336.

Wang, Z., Gerstein, M. and Snyder, M. (2009) ‘RNA-Seq: a revolutionary tool for transcriptomics’, *Nature reviews. Genetics*, 10(1), pp. 57–63.

Wan, Y. *et al.* (2013) ‘Genome-wide mapping of RNA structure using nuclease digestion and high-throughput sequencing’, *Nature protocols*, 8(5), pp. 849–869.

Warner, J. R. (1999) ‘The economics of ribosome biosynthesis in yeast’, *Trends in Biochemical Sciences*, pp. 437–440. doi: 10.1016/s0968-0004(99)01460-7.

Waters, A. J. *et al.* (2017) ‘Natural variation for gene expression responses to abiotic stress in maize’, *The Plant journal: for cell and molecular biology*, 89(4), pp. 706–717.

Watkins, T. N. *et al.* (2018) ‘Comparative transcriptomics of Aspergillus fumigatus strains upon exposure to human airway epithelial cells’, *Microbial genomics*, 4(2). doi: 10.1099/mgen.0.000154.

Weindl, G. *et al.* (2007) ‘Human epithelial cells establish direct antifungal defense through TLR4-mediated signaling’, *The Journal of clinical investigation*, 117(12), pp. 3664–3672.

Weinreb, C. *et al.* (2018) ‘Fundamental limits on dynamic inference from single-cell snapshots’, *Proceedings of the National Academy of Sciences of the United States of America*, 115(10), pp. E2467–E2476.

Wen, K. *et al.* (2016) ‘Critical roles of long noncoding RNAs in Drosophila spermatogenesis’, *Genome research*, 26(9), pp. 1233–1244.

Wertheimer, N. B., Stone, N. and Berman, J. (2016) ‘Ploidy dynamics and evolvability in fungi’, *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 371(1709). doi: 10.1098/rstb.2015.0461.

West, A. P. *et al.* (2011) ‘Mitochondria in innate immune responses’, *Nature Reviews Immunology*, pp. 389–402. doi: 10.1038/nri2975.

West, A. P. *et al.* (2015) ‘Mitochondrial DNA stress primes the antiviral innate immune response’, *Nature*, pp. 553–557. doi: 10.1038/nature14156.

West, A. P. and Shadel, G. S. (2017) ‘Mitochondrial DNA in innate immune responses and inflammatory pathology’, *Nature reviews. Immunology*, 17(6), pp. 363–375.

Westermann, A. J., Barquist, L. and Vogel, J. (2017) ‘Resolving host-pathogen interactions by dual RNA-seq’, *PLoS pathogens*, 13(2), p. e1006033.

Westermann, A. J., Gorski, S. A. and Vogel, J. (2012) ‘Dual RNA-seq of pathogen and host’, *Nature reviews. Microbiology*, 10(9), pp. 618–630.

Westermann, A. J. and Vogel, J. (2018) ‘Host-Pathogen Transcriptomics by Dual RNA-Seq’, *Methods in molecular biology* , 1737, pp. 59–75.

Whaley, S. G. *et al.* (2018) ‘Jjj1 Is a Negative Regulator of Pdr1-Mediated Fluconazole Resistance in’, *mSphere*, 3(1). doi: 10.1128/mSphere.00466-17.

Wickham, H. (2016) ‘Programming with ggplot2’, in *Use R!*, pp. 241–253.

Williams, C. R. *et al.* (2016) ‘Trimming of sequence reads alters RNA-Seq gene expression estimates’, *BMC bioinformatics*, 17, p. 103.

Willis, J. R. *et al.* (2018) ‘Citizen science charts two major “stomatotypes” in the oral microbiome of adolescents and reveals links with habits and drinking water composition’, *Microbiome*. doi: 10.1186/s40168-018-0592-3.

Wilson, D., Citiulo, F. and Hube, B. (2012) ‘Zinc exploitation by pathogenic fungi’, *PLoS pathogens*, 8(12), p. e1003034.

Wilson, D., Naglik, J. R. and Hube, B. (2016) ‘The Missing Link between Candida albicans Hyphal Morphogenesis and Host Cell Damage’, *PLoS pathogens*, 12(10), p. e1005867.

Wimalasena, T. T. *et al.* (2014) ‘Phenotypic characterisation of Saccharomyces spp. yeast for tolerance to stresses encountered during fermentation of lignocellulosic residues to produce bioethanol’, *Microbial cell factories*, 13(1), p. 47.

Win, S. *et al.* (2014) ‘JNK interaction with Sab mediates ER stress induced inhibition of mitochondrial respiration and cell death’, *Cell death & disease*, 5, p. e989.

Wittkopp, P. J., Haerum, B. K. and Clark, A. G. (2004) ‘Evolutionary changes in cis and trans gene regulation’, *Nature*, 430(6995), pp. 85–88.

Wolf, T. *et al.* (2018) ‘Two’s company: studying interspecies relationships with dual RNA-seq’, *Current Opinion in Microbiology*, pp. 7–12. doi: 10.1016/j.mib.2017.09.001.

Wucher, V. *et al.* (2017) ‘FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome’, *Nucleic acids research*, 45(8), p. e57.

Wu, G. *et al.* (2015) ‘Genus-Wide Comparative Genomics of Malassezia Delineates Its Phylogeny, Physiology, and Niche Adaptation on Human Skin’, *PLoS genetics*, 11(11), p. e1005614.

Wu, J. *et al.* (2018) ‘Homoeolog expression bias and expression level dominance in resynthesized allopolyploid Brassica napus’, *BMC genomics*, 19(1), p. 586.

Wu, Y. *et al.* (2016) ‘A Genome-Wide Transcriptional Analysis of Yeast-Hyphal Transition in Candida tropicalis by RNA-Seq’, *PloS one*, 11(11), p. e0166645.

Xie, Y. *et al.* (2014) ‘SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads’, *Bioinformatics* , 30(12), pp. 1660–1666.

Xu, W. *et al.* (2015) ‘Activation and Alliance of Regulatory Pathways in C. albicans during Mammalian Infection’, *PLOS Biology*, p. e1002076. doi: 10.1371/journal.pbio.1002076.

Yang, J.-R. and Zhang, J. (2015) ‘Human long noncoding RNAs are substantially less folded than messenger RNAs’, *Molecular biology and evolution*, 32(4), pp. 970–977.

Yang, Q. *et al.* (2016) ‘Transcriptomics Analysis of Candida albicans Treated with Huanglian Jiedu Decoction Using RNA-seq’, *Evidence-based complementary and alternative medicine: eCAM*, 2016, p. 3198249.

Yang, W. *et al.* (2014) ‘Fungal invasion of epithelial cells’, *Microbiological research*, 169(11), pp. 803–810.

Yang, X. *et al.* (2013) ‘HTQC: a fast quality control toolkit for Illumina sequencing data’, *BMC bioinformatics*, 14, p. 33.

Yano, J. *et al.* (2019) ‘Current patient perspectives of vulvovaginal candidiasis: incidence, symptoms, management and post-treatment outcomes’, *BMC Women’s Health*. doi: 10.1186/s12905-019-0748-8.

Yoo, M.-J., Szadkowski, E. and Wendel, J. F. (2013) ‘Homoeolog expression bias and expression level dominance in allopolyploid cotton’, *Heredity*, 110(2), pp. 171–180.

Yuan, S. and Qin, Z. (2012) ‘Read-mapping using personalized diploid reference genome for RNA sequencing data reduced bias for detecting allele-specific expression’, *IEEE International Conference on Bioinformatics and Biomedicine workshops. IEEE International Conference on Bioinformatics and Biomedicine*, 2012, pp. 718–724.

Yu, G. *et al.* (2012) ‘clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters’, *OMICS: A Journal of Integrative Biology*, pp. 284–287. doi: 10.1089/omi.2011.0118.

Yu, L., Fernandez, S. and Brock, G. (2017) ‘Power analysis for RNA-Seq differential expression studies’, *BMC bioinformatics*, 18(1), p. 234.

Zampetaki, A., Albrecht, A. and Steinhofel, K. (2018) ‘Long Non-coding RNA Structure and Function: Is There a Link?’, *Frontiers in physiology*. Frontiers Media SA, 9. doi: 10.3389/fphys.2018.01201.

Zeilinger, S. *et al.* (2016) ‘Friends or foes? Emerging insights from fungal interactions with plants’, *FEMS Microbiology Reviews*, pp. 182–207. doi: 10.1093/femsre/fuv045.

Zerbino, D. R. *et al.* (2018) ‘Ensembl 2018’, *Nucleic acids research*. Oxford University Press, 46(D1), pp. D754–D761.

Zhang, J.-Z. *et al.* (2014) ‘Mitochondrial DNA induces inflammation and increases TLR9/NF-κB expression in lung tissue’, *International journal of molecular medicine*, 33(4), pp. 817–824.

Zhang, M. *et al.* (2018) ‘Transcriptome analysis reveals hybridization-induced genome shock in an interspecific F1 hybrid from Camellia’, *Genome*, pp. 477–485. doi: 10.1139/gen-2017-0105.

Zhang, N., Park, Y.-D. and Williamson, P. R. (2015) ‘New technology and resources for cryptococcal research’, *Fungal genetics and biology: FG & B*, 78, pp. 99–107.

Zhang, Q. *et al.* (2010) ‘Circulating mitochondrial DAMPs cause inflammatory responses to injury', *Nature*, 464(7285), pp. 104–107.

Zhang, X. *et al.* (2019) 'Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels', *International journal of molecular sciences*, 20(22). doi: 10.3390/ijms20225573.

Zhang, Y. *et al.* (2008) 'Model-based analysis of ChIP-Seq (MACS)', *Genome biology*, 9(9), p. R137.

Zhao, S. *et al.* (2015) 'Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap', *BMC genomics*, 16, p. 675.

Zhao, W. *et al.* (2014) 'Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling', *BMC genomics*, 15, p. 419.

Zhao, X. *et al.* (2018) 'Global identification of Arabidopsis lncRNAs reveals the regulation of MAF4 by a natural antisense RNA', *Nature communications*, 9(1), p. 5056.

Zhao, Y. *et al.* (2016) 'NONCODE 2016: an informative and valuable data source of long non-coding RNAs', *Nucleic acids research*, 44(D1), pp. D203–8.

Zheng, G. X. Y. *et al.* (2017) 'Massively parallel digital transcriptional profiling of single cells', *Nature communications*, 8, p. 14049.

Zhu, W. and Filler, S. G. (2010) 'Interactions of Candida albicans with epithelial cells', *Cellular microbiology*, 12(3), pp. 273–282.

Zhu, Y. Y. *et al.* (2001) 'Reverse Transcriptase Template Switching: A SMARTTM Approach for Full-Length cDNA Library Construction', *BioTechniques*, pp. 892–897. doi: 10.2144/01304pf02.

Zoll, J. *et al.* (2016) 'Next-Generation Sequencing in the Mycology Lab', *Current fungal infection reports*, 10, pp. 37–42.