TESI DOCTORAL UPF 2017

Deep sequencing approaches to investigate the dynamics and evolution of interaction networks of *Candida* pathogens and the human host

Ernst Thuer

Supervisor

Dr. Toni Gabaldón

Comparative Genomics Group
Bioinformatics and Genomics Department
Centre for Genomic Regulation (CRG)

UNIVERSITAT POMPEU FABRA
Departament de Ciències Experimentals i de la Salut

CRG Centre for Genomic Regulation

At each increase of knowledge, as well as on the contrivance of every new tool, human labour becomes abridged

– Charles Babbage

Dedicated to:
A crocodile and a shape

Acknowledgments

The work presented here was funded by the ERCs FP7 framework, in the form of a Marie Sklodowska Curie International Training Network Fellowship.

My thanks go to the network of people surrounding me, however close or far they are. Thanks to the ImResFun with all its special characters. I'm sure we'll meet again and sing old songs one day. Wanna wanna.. Thanks to Karl for the organization.

Thanks to the Gabldón group. Marina and the Greek for their calm wisdom, Laia for the energy, Irene for her tolerance and Cindy for her relentless spirit, and to the newcomers, the hopefully not too apathetic Opathets. Thanks to the colonies, Mr. 'Murica and Mr 'Murcia for the obscure conversations. Also to the wetlab, to kind Ester and the two dancing Stars. Thanks to Toni as well, for letting me find my own way.

And generally everybody for being much better people than I. Best of Luck to all the Gabalgeeks gone by and the Geekaldóns to come, I guess you'll need it.

Thanks to Ljubica, for oh so many reasons.

And Laia, Mind the List.

Abstract / Resum

0.1 Abstract

This thesis describes the application of Next Generation Sequencing, especially RNA sequencing, on the investigation of the pathogenic yeast *Candida parapsilosis*. Pathogenic yeasts of the *Candida* clade are one of the most common hospital derived infections, often with a fatal outcome. We applied modern tools in RNA sequencing based transcriptomics to investigate the unknown, noncoding part of the yeasts transcriptome. The investigation led to a potential noncoding RNA with an important impact on the ability of the yeast to tolerate physiological temperatures and therefore colonize humans. Additionally, using modern transcriptomics, we developed a pipeline that classifies and quantifies allelic expression regulation with limited parental information. The pipeline is specifically designed for the analysis of non-model species. Lastly, in the scope of the thesis, conclusions on the pathogen responses of a human cell line were analyzed and described to evaluate its potential as model system.

0.2 Resum

Aquesta tesi descriu l'aplicació de Next Generation Sequencing, concretament en seqüenciació de RNA, en la investigació del llevat patogen *Candida parapsilosis*. Els llevats patògens del clade *Candida* són els que causen les infeccions més comunes en hospitals, amb un resultat potencialment fatal. Aplicant eines basades en transcriptòmica de RNA no codificant per investigar la part desconeguda d'aquests llevats, hem descobert una seqüencia de RNA no codificant amb un impacte important en la capacitat del llevat per tolerar temperatures fisiològiques i per tant colonitzar els éssers humans. A més, utilitzant la transcriptòmica, hem desenvolupat un programari que classifica i quantifica la regulació de l'expressió al.lèlica i la informació parental limitada. Aquest programari està dissenyat especficament per l'anàlisi d'espècies no-models. Finalment, també es van analitzar les respostes dels patògens en una línia cel.lular humana i es va avaluar el seu potencial com a sistema model.

Preface

This thesis describes the application of modern methods investigating pathogenic yeasts. Projects described cover various aspects of the application of RNA sequencing based transcriptomics, predominantly in the investigation of the pathogenic yeast *Candida parapsilosis*. We investigated novel features of the yeasts cellular mechanisms, developed a new tool for data processing and analyzed the behavior of yeast compared to other pathogens in a human cell line. We hope that the work presented will give insight to more hidden cellular features of the pathogen, enable novel approaches for the study of its transcriptome and improve experimental setup for future analysis.

- **Chapter 1** contains the introduction to the thesis material. It covers pathogenicity of yeasts, RNA sequencing technology, and basic concepts of the applied statistics.
- **Chapter 2** describes the first major project of the thesis. It concerns the investigation of noncoding RNA in the yeast *C. parapsilosis*.
- **Chapter 3** covers a more methodological aspect of the thesis. It describes a developed software approach to the investigation of Allelic Expression in populations of cells.
- **Chapter 4** describes a project investigating the effect of various pathogens on the human cell line THP-1.
- **Chapter 5** discusses the findings of the projects described in the thesis.

Contents

1 Introduction

1.1 General

Biology has developed into a field of quantification. The last few years have seen an incredible increase in the resolution of biological data due to the development of high-throughput technologies such as Next Generation Sequencing (NGS). Yet, like other new technologies, new analytical capabilities bring about new challenges. The recent quantity of data, provided by technologies like NGS, puts into question existing beliefs and creates the need for new approaches to make full use of new observations. Conclusions in Biology have always been gained by empirical research, through observation of changes in phenotypes or patterns. Those observations are a macroscopic measurement of the underlying complex system, which was never completely visible to observers. Such systems present a large degree of homogeneity, as a consequence of scale. An organism of one species will look easily distinguishable from an organism from another. A disease will strike patients within a risk group more likely than others. So the undelying complex systems become simpler if observed from a sufficiently large distance. But in reality, some patients, even among the highest risk group are unaffected, while seemingly arbitrarily chosen individuals succumb to vulnerabilities. Many issues in biology have found their answers in the macroscopic analysis. As, for example, sanitation and antibiotics have removed most diseases from modern society. we are now capable to find the minute details of the underlying complexity. In order to improve upon those earlier approaches. Now it is upon us to find solutions not just for the macroscopic all, but for the individual everything.

The twenty first century has seen an unprecedented speed of technological

advances. An important technical advance in biology, related to the work presented here, was the introduction of sequencing technologies. Sequencing refers to the decoding of the DNA sequence, the exact ordering of nucleotides in a DNA molecule. In the beginning sequencing was very limited in scope and was commonly applied only to individual genes. The sequencing field moved into the spotlight of the wider public with the Human Genome Project (Hood and Rowen, 2013), but consisted of several steps of technological advancement. The technology that started the trend was developed by Frederick Sanger et alii, described in 1977 (Sanger et al., 1977). The first commercial development was produced by Applied biosystems, enabling standardized procedures for biology. The twenty first century saw several leaps in development. Solexa released its first sequencer in 2006. The company was bought by Illumina in 2007. The new high throughput approach of this system allowed Illumina to establish itself at the basis of most NGS analysis. As compared to Sanger sequencing which, e.g on a commercial 96 capillary instrument produced 6 Mega bases per day (Kircher and Kelso, 2010), the Illumina system gave researchers the ability to sequence 1 giga bases in a single run. Nowadays, sequencing throughputs in the tera-base range are common. This represents an increase by a factor of 1000 in only 10 years.

The field of Bioinformatics is not old by any measure. The need to combine biology with the power of informatics is relatively new, at least in the current scale. Considering storage necessities, NGS combined is the largest data producer worldwide, with an annual need for storage increase between 2 and 40 Exabytes predicted by 2025 (Stephens et al., 2015). Raw data generation volume is only overshadowed by astronomy. An online database tool, OmicsMaps [omicsmaps.com] traces the distribution of individual sequencing machines worldwide, although due to the rapid speed of development, even an online community gets outdated quickly. A paper from 2015 listed 2,500 sequencing machines (Stephens et al., 2015) on omicsmaps. There are currently 7,389 total sequencing machines listed in this database, distributed in 1,027 sequencing centres. With the amount of sequencers predicted to increase sharply over the coming decades. Unlike other large producers of data, biological information is very heterogeneous (Stephens et al., 2015). Derived from biological systems, the data is

inherently noisy and lacks the reproducibility observed for example in physics. Individual sequences are obtained from groups of organisms under potentially heterogeneous conditions and data is also created by different researchers, introducing varying levels of human error. Computers had to be used early on to handle the amount of data obtained even from earlier sequencing approaches such as Sanger sequencing. The originally sporadic necessity has developed into a field of its own (Hagen, 2000). Originally, many bioinformatics algorithms were borrowed from other fields such as sociology, physics or economics. Over the last few years dedicated approaches yielded algorithms better suited to handle large scale data, and the analytical approaches have become more standardized. The increase of computational speed was predicted by Moore's law (Muir et al., 2016). This law states that the number of transistors in a dense integrated circuit double roughly every year (Stephens et al., 2015). This advance has seen its limitations over the last years due to lack of space and cooling capabilities for additional transistors. Alternative methods, such as multi core processing take the place of denser transistor assembly, yet with limited increases. Slowing down the advance of computational power. An important problem in bioinformatics is that the speed at which new data is generated has outpaced the increase in computational speed. As a result there is a pressing need to develop new algorithms that are able to analyze larger amounts of data with relatively smaller computational resources. And even as the algorithms handle the increase of data, the simple need to store the data becomes a burden on many projects. Predictions estimate the need for storage of data to increase between 2 and 40 Exabytes per year by 2025. Data storage and curation has become a challenge by itself, and will likely be a major concern in the coming future. In addition a gap of knowledge has appeared, and the shortage of experienced data analysts cause problems in many advanced research projects (Sboner et al., 2011).

1.2 Biology of pathogenic Fungi

Human fungal pathogens are add odds with classical definitions of pathogenicity in that, for them, the immunological status of the host is a cen-

tral determinant of the outcome of infection (Casadevall and Pirofski, 1999). Over the last decades, medicine and sanitation have made great advances in the treatment and prevention of most human diseases. Human lifespan is increasing, and medical care cures most types of even the deadliest diseases. But with the increase in care, new niches have been formed that allow new opportunistic pathogens to take over. For example in the treatment of cancer, in which actively dividing cells have to be depleted, immune cells are often targeted leading to a weakening of the immune system and, in turn, paving the way for the onset of life threatening infections. Similarly, in diseases like Acquired Immune Deficiency Syndrome (AIDS), or during the treatment of autoimmune diseases similar conditions are created. In immunocompromised patients, opportunistic pathogenic yeasts are becoming a cause of growing medical concern. Often harmless commensals in most humans, these opportunistic pathogens can cause life threatening infections that are difficult to diagnose on time and hard to treat. Amongst those opportunists are a variety of species from genera such as *Aspergillus*, *Cryptococcus* as well as several species of *Candida*. *Candida albicans*, *C. glabrata* or *C. parapsilosis*, which belong to a phylogenetically diverse clade, *Saccharomycetaceae*, that also includes *Saccharomyces cerevisiae* (Gabaldon et al., 2016). The molecular basis of the infection mechanisms in the different species remain poorly understood, and it is as yet unknown to what extent the host-pathogen interactions for the different species differ or resemble each other. Recent developments in high-throughput sequencing techniques enable studying the transcriptional behavior of host and pathogen simultaneously and with unprecedented resolution.

1.2.1 *Candida* the yeast

Yeasts are unicellular fungi. Like all fungi they are eukaryotes, and therefore larger and more complex than prokaryotic organisms, with whom they compete for resources. A suggested review by (Stajich et al., 2009) goes into more details. Yeasts have remarkable cellular properties, that helped them adapt to an almost prokaryotic lifestyle. They have short generation times, withstand wide ranges of temperatures and are able to survive on very simple nourishment. Among yeasts, the clade most closely associated to

humans are the saccharomycotina. Saccharomycotina, although considered one clade, encompass species as distant as humans and fish (Dujon et al., 2004). *Saccharomyces cerevisiae* is one of the best studied model organisms in biology. It is commonly used in a variety of biotechnological processes such as food processing and others. Yet, given the right conditions, *S. cerevisiae* can be pathogenic, causing potentially life threatening systemic infections (Murphy and Kavanagh, 1999). Most notably in immunosuppressed individuals. Many other species of saccharomycotina have evolved this ability to cause infections in humans in various ways. Their commensal pathogenic behaviour extends their range of environment to mammalian hosts. Most interactions of saccharomycotina and mammals appear to be commensal, meaning not to the detriment of either the host or the yeast. Human-associated yeasts populate mostly mucosal surfaces, intestines and genital tracts. But the variability of yeasts enabled the individual species to adapt to their mammalian hosts in different ways. Some species, like *Candida albicans*, have adapted to a life of almost obligatory commensalism. Tolerating the human body temperature and avoiding the activation of a defensive immune system, while integrating into the microbiome. However, this commensal balance is not perfectly stable. In a weakened host, *C. albicans* may switch its morphology and cellular behavior to an invasive one. This species has adapted a series of mechanisms to counter human defenses and that can be harmful in a weakened hosts. Potentially killing a weakened host in a matter of days.

There are around 200 species of *Candida* described, around 20 of which are considered pathogenic. *Candida* are often diploid with some notable exceptions (Gabaldon et al., 2016). They contain a genome of only about 12 to 14 million base pairs. Up to 75 % of their genome is considered to be protein coding. They also share an altered genetic code, in which the CUG codon is translated as leucine as opposed to serine (Santos MA, 1995). Introns are not common, with only about 10 % of the genes showing multiple exons. There are 499 existing intron annotations on 6,218 genes in the *C. albicans* SC5314 strain. *Candida* and other yeasts have evolved mechanisms to increase genetic variability. They maintain a very fluid genome, which shows high variability across related species. Several strains

are hybrids offsprings of related species (Pryszcz et al., 2015). Individual strains in this diverse clade have developed mechanisms to generate more genetic diversity via diverse types of sexual cycles. Mating is based around two mating types, a and alpha, that are encoded in the Mating Type Locus (MTL). *Candida* species were originally defined as species that can form pseudohyphae or true hyphae but lacked any form of sexual reproduction (Reedy et al., 2009). Observations in the early 2000s have changed this definition, with discoveries that e.g *C. lusitaniae* has a defined sexual cycle presumed to include meiosis based on its ability to produce spores (Reedy et al., 2009). The mechanisms in other species were less clear, *C. albicans* for example, presents orthologs for most genes necessary for sexual reproduction compared to *S. cerevisiae*, yet has never been observed undergoing mating. Recent studies discovered a parasexual reproduction cycle in *C. albicans*, and potentially *C. dubliniensis*. This cycle involves mating, recombination, and genome reduction but with no recognized meiosis. In *C. albicans* parasexual mating requires a shift in morphology, only opaque colonies are known to mate (Bennett, 2015).Other species, like *C. parapsilosis* have most likely entirely lost the ability to mate (Sai et al., 2011). In this species, only MTLa/a is present and the MTLa1 gene is a pseudogene, which suggests that the MTL locus might be degenerating. *C. Parapsilosis* has two closely related yeasts, *C. metapsilosis* and *C. orthopsilosis*. In *C. orthopsilosis*, but not *C. metapsilosis* there is a mixture of mating types, orthologous to *C. albicans*, suggesting the presence of an extant sexual cycle (Sai et al., 2011). This dynamic between closely related species exemplifies the fluidity of the *candida* genome. A more extensive review on mating in human fungal pathogens can be found at (Ene et al., 2014). *Candida albicans* is the best studied and most pathogenic of all *Candida* species. The current state of *C. albicans* genome annotation comprises 4,422 predicted ORFs (71.12%) 1,644 verified ORFs, (26.44%) and 152 dubious ORFs (2.44%) (Arnaud et al., 2005). This species has evolved to become an obligatory human commensal, and is found in the human intestinal track of 30% to 70% of the population (Hoffmann et al., 2013). The species studied in most projects described here is *C. parapsilosis*, which shows a comparable genome structure, protein count and gene density. *Candida* species colonize a variety of environments. Unlike *C. albicans*, *C. parapsilosis* for example is not an

obligatory commensal, but occurs naturally in soil or marine samples, yet also colonizes mammalian hosts. The ability of different *candida* species to colonize various environments such as soil or non human hosts is still being investigated. Recently *C. dubliniensis* for example was originally considered to be restricted to the human microbiome (Sullivan et al., 2005), but has since been found colonizing the seabird tick *Ixodes uriae* (Nunn et al., 2007). Environmental occurrence of different potentially pathogenic species may be a consideration in connection to global warming fostering additional temperature adaptations in otherwise environmental strains, potentially increasing their pathogenicity.

The distribution of *Candida* strains varies by country or region, and can change drastically in different hospital environments (e.g (Singh and Parija, 2012)). The usually most common species is *C. albicans*, followed by either *C. glabrata* in northern europe and the U.S and *C. parapsilosis* in e.g Spain or Brazil (Guinea, 2014). Across hospitals the distribution varies again, with individual institutions maintaining their respective prevalent strains. This is an important problem in the treatment of *Candida* species. *C. dubliniensis* is more resistant to fluconazole for example, while *C. parapsilosis* is more susceptible. Knowledge of potential presence of certain strains could influence modes of preventive treatment.

There are mixed reports on the effect of *Candida* on the healthy human body. The presence of different *Candida* species can be used as clinical markers in dosage of cholesterol medications like statins, that affect the fungal counterpart ergosterol (Wikhe et al., 2007). Little is known about the interactions between *Candida* species and the healthy human microbiome, but due to its widespread presence and the lack of active removal by healthy hosts, it seems that *Candida albicans* is not an unwelcome parasite. However, *Candida* cause occasional problems even in healthy hosts. The most prevalent clinical condition is vulvovaginalcandidiasis (VVC), with more than 10 million cases annually (Linhares et al., 2001). Up to 90% of sporadic, uncomplicated cases of VVC are caused by the species *C. albicans*, followed by 5% to 11% *C. glabrata* and between 4 and 8 % *C. Parapsilosis* (Nyirjesy et al., 2005) and are treated with single-azole medication. Historically, species identification has been omitted, due to the relative susceptibility of

all treated strains to the admitted antifungals. An important observation was that the amount of *candida* cells were not correlated to the occurrence of clinical symptoms (Linhares et al., 2001), unlike pathogenic bacteria or viruses, quantity does not cause an infection.

Healthy *Candida* commensalism, even if not proven to be beneficial, is benign. Potential danger to patients occurs when the yeast senses a physiological change in the human host. A number of criteria seem to have an effect on *Candida* behavior. An infection caused by *Candida* is termed candidiasis or candidosis (Sardi et al., 2013). Those infections may present a variety of clinical symptoms (Sardi et al., 2013). An important consideration when studying candidemia is that the most common way of transmission for candidiasis is endogenous. This occurs when species that constitute the microbiota of various anatomical sites under conditions of host weakness may behave as opportunistic pathogens (Colombo and Guimarães, 2003). The most important trigger condition for endogenous candidiasis in humans is acute neutropenia, and a depletion of T- helper cells. If *Candida* senses an immunological weakness, its behavior shifts drastically within a few hours. In cultures, this can be observed by a morphological shift visible in plated colonies, called white - opaque switching. *Candida* cells grown on rich medium form white colonies. This macroscopic shift is a result of morphological changes, *C. albicans* is able to grow as ovoid budding yeasts, pseudohyphae or true hyphae (Berman and Sudbery, 2002). Additionally it has the ability to form chlamydospores of unknown function (Citiulo et al., 2009). The observed dimorphism is an important feature in biofilm formation and mating (Lohse and Johnson, 2010), (Lockhart et al., 2002). *C. albicans* maintains the white state during infectious behaviour, e.g at 37°C, but requires the opaque state to undergo mating (Lohse and Johnson, 2010). It has more recently been discovered, that the opaque state is stable at 37°C under anaerobic conditions (Dumitru et al., 2007). Shifting from opaque to white is characterized by changes in morphology, but especially in the formation of hyphae by the *C. albicans* cells. Only *C. dubliniensis* is capable of the same white opaque switching. *C. dubliniensis* has even been reported to be able to mate with *C. albicans* (Pujol et al., 2004). Hyphae start their development with the formation of spitzenkorper. Those spitzenkorper are

being elongated, extending the cell shape, and show, in *C. albicans* a unique expression profile compared to pseudohyphae (Crampin, 2005). Hyphae maintain a rigid cell wall and extend the physical range of the fungus. But they also act as invasion organs. *Candida* ingested by macrophages use hyphae to cross the membrane of the immune cells and break free, destroying the immune cells in the process. Hyphae are also used to transgress human epithelial cell layers, permeating into the bloodstream. Another important feature, that is widespread also among *C. albicans* sister species is the ability for form biofilms. This is especially important in the hospital environment. Biofilm formation has been shown to be triggered by pheromone secretion of fungi (Hawser and Douglas, 1994; Chandra et al., 2001). They are more resistant to stresses and limit the efficiency of drugs due to lower accessibility of individual cells. Biofilms are hard to combat and can establish themselves on clinical devices such as catheters. Once established they can provide a constant source of new fungal cells entering the host.

1.2.2 Clinical importance

Candida species are the 4th most common source of nosocomial (hospital-derived) infections. Candidiasis shows a prevalence of up to 60 per 100.000 hospital admissions (Wisplinghoff et al., 2004) (Brown et al., 2014). Systemic infections have a mortality rate of more than 40% for systemic candidiasis (Amorim-Vaz et al., 2015; Falagas et al., 2006). Medical research has improved patient care drastically over the last decades. With more patients surviving an immunocompromised state for extended periods of time, *Candida* have found a new niche.

As discussed above, pathogenic yeast are opportunistic pathogens, making pre-existing conditions such as immunosuppression an obligatory condition for their pathogenicity. Different species of *Candida* have different preferred host spectra. E.g *C. parapsilosis* is prevalent in neonates (Chow et al., 2012), while *C. glabrata* is more common among the elderly. (Fidel et al., 1999) *Candida* can cause life threatening systemic infections. Infections are most common among immunosuppressed individuals, patients with acute renal malfunctions, cancer or transplant recipients, as well as AIDS patients

in advanced stages. Added to this is that the most common antifungal treatments have serious side effects, and detection and classification of the right *Candida* species is slow, and require special equipment. Specialized diagnostic centers use mass spectrometry to rapidly detect the species, and recent developments point to a potential use of this technology in detecting antifungal resistant strains (Saracli et al., 2015).

The human intestinal tract potentially provides rich media for the yeasts to grow, assuming it can adapt to the adverse conditions like body temperature and the complex microbiome (Netea et al., 2015). 37° C are suboptimal growth conditions for most species of fungi. In recent studies, up to 60% of healthy humans are positively tested for *Candida* species (Hoffmann et al., 2013). Pathogenic behaviour requires an additional adaptation to environments within the bloodstream or other organs. Certain environments in humans, like the bloodstream show a tightly constrained limitation of some nutrients or elements, like iron, which is an important component of various biologically important enzymes. Iron fulfills similar functions in yeasts, and the host can reduce its availability using iron-chelating proteins in order to limit its accessibility to yeasts (Knight et al., 2005). Specific species, that manage to adapt to the environmental conditions face a complex immune system that will be activated by a variety of molecular patterns. For the most part the yeasts do not seem to damage the host, and the immune system has adapted to accept their presence under such circumstances. The immune response triggered by the pathogen is complex, and out of scope for this introduction. Reviews and detailed descriptions are available (Brown et al., 2014; Gow et al., 2011) To adapt to this set of responses, fungal cells posess several virulence mechanisms. Such mechanisms can be quite distinct within the clade, and will be discussed for some specific *Candida* species in the following section.

1.2.3 Mechanisms of Pathogenicity

Candida parapsilosis

Most work presented here pertains to a related species to *Candida albicans*, *C. parapsilosis*. Originally thought to be a single species, it was recently split into three distinct species, *Candida parapsilosis* sensu stricto, *C. orthopsilosis* and *C. metapsilosis*. *C. parapsilosis* and its close relatives differ in some fundamental mechanisms of survival and in many ways in their approach towards the human host. They are considered to be less prevalent in infecting humans than *C. albicans*. The potential pathogenicity is highest for *C. parapsilosis*, followed by *C. orthopsilosis* and the relatively weak pathogen *C. metapsilosis* with percentage of relative incidences of 90, 8 and 2 % respectively observed in VVC (Lockhart et al., 2008; Bonfietti et al., 2012). All three species have a global distribution and can exist both as commensals within the human microbiome, and as environmentals isolated from a variety of niches such as the hands of healthcare workers (Sabino et al., 2015) but also ranging from oceanic to arboreal. Considering clinical isolates, *C. orthopsilosis* and *C. metapsilosis* show a very low prevalence (1 %) in most bloodstream isolate collections surveyed. Additionally they tend to respond well to antifungal drugs (Falagas et al., 2010).

Clinical importance in *C. parapsilosis* derives at least partly from its shift in host spectrum compared to *C. albicans*. This species shows higher prevalence in neonates (Neu et al., 2009; Roilides et al., 2004). Another major factors of *C. parapsilosis* pathogenicity are its increased tendency to adhere to plastics, like catheters in hospitals (Branchini et al., 1994), and its ability to form biofilms, as it shows a high prevalence of forming biofilms on plastic surfaces comparable to that of *C. albicans* (Kuhn et al., 2002). *Candida albicans* hyphae formation is an obvious morphological switch, and is the most easily observable difference to *C. parapsilosis*. *C. parapsilosis* is not capable of forming true hyphae, a trait considered important for *C. albicans*'s pathogenesis. It has been shown that *C. parapsilosis* only forms pseudohyphae (Berman and Sudbery, 2002). However it still manages to be, albeit to a lesser degree, pathogenic. Pseudohyphae resemble elongated, ellipsoid yeast cells. They remain attached to one another at the constricted

septation site and grow in a branching pattern that is thought to facilitate foraging for nutrients away from the parental cell and colony. True hyphal cells on the other hand, are long and highly polarized, with parallel sides and no obvious constrictions between cells. The effect of pseudohyphae on *C. parapsilosis* pathogenicity has been the focus of recent research. Given the lack of hyphae formation, other mechanisms have to retain *C. parapsilosis* potential as a pathogen. Under investigation are pathways concerning the production of prostaglandins (Grózer et al., 2015), biofilm formation (Singh and Parija, 2012) and *C. parapsilosis* interaction with macrophages (Tóth et al., 2014).

Another important consideration is the lack of even a parasexual life cycle, as in other *Candida* species (Bennett, 2015). In *C. albicans* morphology and reproduction are linked. Only the opaque morphologies is able to undergo the reproductive cycle (Dumitru et al., 2007). So a lack of morphological switching and parasexual cycle may not be independent occurrences.

Studies on the distribution of candidemia showed an increase in *C. parapsilosis* cases over the years. In clinical observations in Spain and Brazil, *C. Parapsilosis* has overtaken *C. glabrata* in overall numbers of infected patients (Guinea, 2014). Predisposition of an infection with *C. parapsilosis* is dependant on the environment, with most clinically relevant cases observed in hospitals, with stationary patients. In some hospitals *C. parapsilosis* infections even outranks *Candida* albicans (Singh and Parija, 2012).

Our knowledge of *C. parapsilosis*, even though a relatively common human commensal or pathogen, is far from complete. Although recent years have seen increases in coverage of genomic studies and new investigations into behavior of pathogenicity, many underlying mechanisms remain to be studied.

Candida albicans

Candida albicans is the most intensely studied *Candida* species. It is also the most prevalent. Up to 50-60 % of a healthy human population carry this species in their normal microbiome (Hoffmann et al., 2013). *Candida albicans*

causes most of the systemic *Candida* infections, up to 90% of (Hoffmann et al., 2013) Genital / Vulvovaginal Candidiasis (VVC) and around 50% of systemic candidiasis (Nur, 2014). It has a variety of mechanisms directly associated to its pathogenic potential. The most intensely studied is the above mentioned switching of morphology. This shift in morphology is one of the core mechanisms of pathogenicity, together with its pseudo sexual cycle.

Recent discoveries have broadened our understanding of *Candida* pathogenesis. A recent study published by a collaborator from the same network shows the existence of a *Candida* lysin, a peptide toxin secreted by *C. albicans*. The lysin is the product of a KEX2 digestion of the ECE1 protein, and seems to be an essential factor for mucosal invasion (Moyes et al., 2016). This added a new layer to an already complex network of interactions.

Overall, antifungal resistance is a less important consideration of *Candida albicans*. Known isolates seem to adapt slowly to the available drugs. Current isolates are susceptible to a range of antifungals such as fluconazole and Amphotericin B, mechanisms for their functionality are mentioned below.

Candida glabrata

Candida glabrata is evolutionarily distant from other *Candida* species. It differs greatly from *Candida albicans* in its mechanisms of pathogenicity, such as hyphae formation and peptide secretion, but does form biofilms and pseudohyphae (Kaur et al., 2005). Compared to other *Candida*, *C. glabrata* is much more readily adapting to antifungals, and is innately resistant to azole based antifungals. This resistance seems to be mediated by overexpression of several ABC transporters (Miyazaki et al., 1998). *C. glabratas* recent increase in prevalence as a human pathogen is likely linked to its drug resistance. The preemptive administration of azole antifungals creates a niche in patients. Multiple antifungal resistances in *C. glabrata*, especially in absence of the development of new antifungals are becoming a major threat. Evolutionarily, *C. glabrata* is more closely related to *S. cerevisiae*, which can also be pathogenic (Murphy and Kavanagh, 1999). *C. glabrata* was assumed to show a non dimorphic blastoconidia morphology and has

a haploid genome. Yet recent studies show morphological switches for *C. glabrata* with several distinct phenotypes observed (Lachke et al., 2000, 2002). Additional remarkable traits of *Candida glabrata* include its short generation time, compared to other fungi. It is capable of overgrowing the immune system. The generation time, under the right conditions can be as low as 1.02 hours at 37°C (Roetzer et al., 2011). *C. glabrata* is not an obligatory commensal. It can be found environmental, but is often associated to different body sites, forming part of the human microbiome.

Other *Candida* species

The candida clade comprises around 200 species. Definitions of the clade have changed with the advances of genomic sequencing. The earlier morphological definitions did not describe the true variability of the clade. Only around 15 candida species are of clinical importance (Turner and Butler, 2014). Apart from the ones mentioned previously, only *C. dubliniensis*, *Lodderomyces elongisporus* and *C. tropicalis* compose a mentionable portion of clinical isolates.

1.2.4 Current treatments

The treatment of *Candida* infections relies heavily in risk group assessment. *Candida*, as a commensal is widely spread, but as a pathogen, targets only a limited range of potential patients. Immunosuppressed patients often receive preventive medication in order to halt infective progression. Such preemptive treatment is necessary, since detection of an infection is often slow, and infections may lead to acute septic shock (Delaloye and Calandra, 2014), with mortality rates around 30 - 60% (Delaloye and Calandra, 2014; Hirano et al., 2015). Preemptive treatment is an important factor for development of antifungal resistances, especially in *C. glabrata*.

A few antifungal drugs are available to treat systemic candidiasis. As in antibiotics available for prokaryotes, antifungals target essential but distinct cellular mechanisms in yeasts. An important consideration is that yeasts are eukaryotic organisms, as are their hosts so the potentially targetable

mechanisms are significantly more limited as compared to antibiotics for prokaryotes. The drugs should avoid target mechanisms that are closely related to their counterparts in humans. Thus, primary targets of antifungal drugs are ergosterol metabolic pathways, which are absent from human cells. Antifungal drugs have a variety of adverse side effects, particularly considering that they commonly administered to already weakened patients. The most important antifungals are Fluconazole / Azoles and Amphotericin B and echinocandins, such as caspofungin. Amphotericin B induces potassium permeability, thereby destabilizing the cell wall. There is a documented effect on cholesterol containing cell walls (e.g in humans), but at exposure to much higher concentration (e.g (Bolard et al., 1991)).

Fluconazole acts on a different path of ergosterol maintenance inside a fungal cell, like other azoles, it interrupts the conversion of lanosterol to ergosterol via binding to fungal cytochrome P-45. This disrupts the fungal membranes (Zervos and Meunier, 1993; Pasko et al., 1990).

Echinocandins, such as caspofungin, are large lipopeptide that inhibit the synthesis of -(1,3)-glucan. They are effective against *Aspergillus* and *Candidas*, but in clinical concentrations not usable against *Cryptococcus*. *C. parapsilosis* also shows an increased Minimum Inhibitory Concentration towards echinocandins (Denning, 2003).

1.2.5 Model organisms in *Candida* research

The gold standard of animal models used in *Candida* pathogenesis research is the mouse, both transgenic and wild type (LePage and Conlon, 2007; Yano and Fidel, Jr., 2011; Conti et al., 2014). Most models have been established for *C. albicans*, but are also used in the analysis of *C. parapsilosis* and *C. glabrata*. An important consideration amongst various *Candida* species is their host distribution. While all of the above mentioned yeasts colonize humans, neither *C. albicans* nor *C. parapsilosis* colonize mice under normal conditions. This is an important consideration in the analysis of experimental data from mouse models. Since mice are not a real host for the species, their response and management of fungal burden may differ significantly from humans. Mice have to be severely immunocompromised to maintain a systemic

infection (Jacobsen, 2014). Fungal burden in humans is carried by the spleen, while in mice the burden is shifted towards the kidneys (described e.g in (Szilagyi et al., 2012)). A very different environment. Yet mice are the default model for studies in fungal burden. Several considerations complicate interspecies comparative analysis. Intravenous injection of C. albicans is lethal, with the cause of death attributed to fungal sepsis (Conti et al., 2014), but e.g *C. parapsilosis* is not fatal, even in heavily immunosuppressed mice. Mouse experiments are costly, logistically challenging, time-consuming and ethically delicate. Over recent years, another animal model has gained attention in the field. *Galleria mellonella* has been established as an invertebrate substitute host for different *Candida* infection models. (e.g (Jacobsen, 2014; Cotter et al., 2000)). *G. mellonella*, also known as honeycomb moth, is a fast growing insect. Its usage reduces the time and cost of studies in pathogenicity, as well as presenting fewer ethical considerations. Its larvae are used as infection models for yeasts and other pathogens like *Cryptococcus spp* (Jacobsen, 2014). This moth is considered a good substitute for mammalian host, with its innate immune system considered closely resembling that of vertebrates. The most prominent problems with this model are the lack of an acquired immunity system, which is missing in all insects, and the difference in optimal temperature. One benefit of *G. mellonella* is its temperature resistance, models are established at 30° C and 37° Celsius (Fuchs et al., 2010), yet its optimal growth temperature is 30°C. Although the overlap of observed virulence phenotypes between mouse and *Galleria* models has been reported to be as low as 25%, *G. mellonella* can be used as a valid model e.g for screening mutants with important defects in host infection (Amorim-Vaz et al., 2015). Both *G. mellonella* and mouse models were used during the analysis of phenotypes of long noncoding RNAs in the project in chapter 2. Due to the ease of handling and experimental setup, the *G. mellonella* model was used widely on all five deletion mutants. This provided an initial insight into the potential impact of the deleted transcripts. Only one of our potential noncoding RNA phenotypes was further analyzed in a mouse model. Important distinctions between the models from the perspective of produced data comes from organ resolution. *G. mellonella* gives a single mortality response while the mouse model provides more nuanced organ specific responses.

Another important model is the use of human cell lines. The application of cell lines as a reproduction of a complex process like infection is difficult due to the mixture of cell types affected during infection. Most cell lines used are secondary cell lines meaning cancer derived. This potentially influences the accuracy of their response, due to abnormal behaviour of cancer cells. Most common cell types include epidermal cells to simulate mucosal surfaces, as well as Immune cells to simulate systemic infections. We describe an investigation of one commonly used cell lines THP-1 comparing its behaviour in response to various pathogens, viral, prokaryotic and eukaryotic in Chapter 4. An important distinction of single cell lines is the relative ease of transcriptomics investigation, with the drawback of an unnatural environment.

1.2.6 ImResFun network and collaborations

The projects described here have been made in the context of a Marie Curie International Training Network (ITN). This network, called ImResFun, combined ten academic and three private institutions across Europe to address a common research theme. The project aimed at a combined investigation of host-pathogen interactions in *Candida* pathogenesis. Several projects were initiated with collaborators from within this network. The CRG, was the only academic partner with a computational focus. Complemented with the commercial CLC bio, by now Qiagen Bioinformatics. Most partners in the network are experimental research groups, and the individual members experimental scientists. All partners are leading experts in their respective fields that have driven the progress in the field of *Candida* research. The range of research thematics was quite broad. But the overall focus followed the pattern of a deep investigation of the complex pathogenic systems of *Candida* species. With groups working on a variety of fields. Amongst the intitiated projects were the establishment of mixed tissue models to simulate human mucosal surfaces during the *Candida* infection, the discovery of *Candida* lysin, and others. The CRGs research focus in the course of the network was in part advisory, providing consultancy on experimental and analytical designs as well as carrying out analysis in collaboration with the network. The other, and more extensive, part concerned the deepening of our under-

standing of cellular features in *Candida* yeasts. Extending existing annotations into noncoding genome features. Additionally, projects were initiated to increase our insight using modern transcriptomics data, and to contribute to methodological improvements into those complex analysis.

1.3 Extending Biology

1.3.1 Noncoding Features in Yeast

The analysis of genomic features has improved in parallel to the development of of new methods and technologies. With the advance of genomics, and the closer investigation of the *C-value enigma* (Gregory, 2007) novel transcriptional behaviour was proposed. The amount of protein-coding genes in Humans was not considered by some as sufficient to allow their complexity. This forced researchers to reevaluate expectations, that translated genes account for all cellular diversity. Early experiments in cellular expression using tiling microarrays and sequencing established that protein coding regions were only a part of the expressed and regulated transcriptome. Noncoding functional RNA is still a subject of debate in the field of biology. A series of experiments done on hybridization of cDNA via genomic arrays (e.g (Bertone et al., 2004)) or library sequencing (Okazaki et al., 2002) (Carninci et al., 2006) have provided a comprehensive transcriptional landscape within mammalian cells. Recently the ENCODE project provided a more comprehensive effort in surveying transcription in human HeLa cells (ENCODEProjectConsortium, 2007) (HumanGenomeProjectConsortium, 2012) (Derrien et al., 2012) (Djebali S, 2012). ENCODE provided proof of pervasive transcription in the human cell. It seems that biological systems are complex, and in flux, and classical axioms, such as the necessity for all functional elements in a cell to be converted to amino acids have been revisited in the light of new data. A variety of features are currently attributed to such expressed noncoding transcripts, from chromatin regulation and genetic imprinting to protein expression regulation (Wilusz et al., 2009) (Schmitz et al., 2016). Comparable with the early stages of protein coding gene analysis, the study of noncoding RNA is only starting to get an insight into previously

hidden features of cellular organization. Yet the true criteria that define a functional noncoding transcript are still unclear. Crude cutoffs are currently in place, to distinguish types of noncoding RNAs. A length cutoff of 200 nt is used to distinguish lncRNAs from short noncoding RNAs like the 21 to 35 nt microRNAs, Piwi-interacting RNAs (piRNAs), and small-interfering RNAs (siRNAs) like RasiRNAs (e.g. in a review by (Carmi, 2006)). Even the definition of long noncoding RNA has become heterogeneous, several types of long noncoding RNAs have been described. Antisense transcripts, pseudogenes, or long intergenic noncoding RNAs (Wilusz et al., 2009).

In humans many cellular functions of long noncoding RNAs were studied in cancer sets, and are therefore attributed to related clinical processes. The best studied examples involve chromatin modification, X chromosome silencing *Xist* (Pontier and Gribnau, 2011), as well as cellular proliferation *HOTAIR* / *MALAT1* (Cai et al., 2014; Gutschner et al., 2013).

The project described in chapter 2 is concerned with the analysis of long noncoding RNA in the above mentioned species of *C. parapsilosis*. Little is known about noncoding RNA in yeast species. The most intensely studied species, *Saccharomyces cerevisiae* has several annotated pseudogenes, antisense transcripts, etc. and currently 17 annotated long noncoding RNAs (Cherry et al., 2012). Two lncRNAs have some functional annotation, with the most deeply studied transcript showing activity in gametogenesis (Yamashita et al., 2016) and involvement in the function of the GAL10 cluster (Houseley et al., 2008). Although a few annotations exist in *C. albicans*, based on tiling arrays (Sellam et al., 2010) no lncRNAs were so far functionally investigated for either *C. albicans* or *C. parapsilosis*. A recent publication on noncoding RNAs focuses on the annotation of short nucleolar RNAs (Donovan et al., 2016). An important consideration when analyzing lncRNAs in yeast is the very different genomics structure. Yeast genomes are very dense, with up to 75% of their sequence being protein coding. Alternative splicing events are also very rare, with only 1 in 10 genes containing intronic regions. So ultimately, conclusions gained from noncoding RNAs in humans may only have limited similarities in yeasts and vice versa. Chapter 2, deals with the analysis of long noncoding RNA in yeasts.

1.3.2 Allele Specific Expression

As with noncoding features mentioned previously, NGS usage in transcriptomics has revealed several features of genomic regulation at the transcript level. There is, at least in Mammals, an interplay between long noncoding RNA and Allele Specific Expression (ASE). The noncoding RNAs *Xist* and its antagonizer *TSIX* regulate the inactivation of the second X chromosomes in females (McHugh et al., 2015) (Pontier and Gribnau, 2011). The availability of the sequence information for sequenced transcript, enabled the quantification of Allele Specific Expression (ASE), that is the finding that different alleles of a given locus can be expressed at different levels and even regulated differently.

ASE can potentially occur on any locus of a polyploid organism. Yet is only detectable if the transcribed sequences differ between the two alleles. Differential Expression of individual alleles has shown to be present in a variety of conditions, most notably in embryonic development (Szabo and Mann, 1995; Szabo et al., 2002; Springer and Stupar, 2007) or diseases e.g of the heart muscle (Sigurdsson et al., 2016). The most common analytically relevant difference between alleles are Single nucleotide Polymorphisms (SNP). In some yeast species, sexual and pseudosexual life cycles are intoducing haplotype diversity within a genome. Additionally hybrid species, resulting from crosses between two distinct parental species, can be found. The overall divergence retains the haploid SNPs of the parentals on single Alleles, leaving sets of biallelic SNPs. Considering protein coding genes, those SNPs can be either sense or missense concerning the amino acid sequence. In certain cases those missense SNPs may result in functional implications of ASE, but in any case both sense and missense SNPs can be used to quantify the extent of ASE. So far, software implementations for the detection and quantification of ASE in nonmodel species, i.e with a missing phased reference genome, are lacking. There have been some statistical frameworks proposed, albeit these are targeting error distributions following single cell transcriptomics (Jiang et al., 2017; Sigurdsson et al., 2016), or analysis in species with phasing knowledge of reference SNP distribution. The project described in Chapter 3 deals with the analysis of Allele Specific

Expression from populations of the hybrid yeast *C. orthospsilosis*. To enable the accurate quantification and classification of the expression data, we implemented a more robust quantification software. Based on the SNP wise analysis of gene expression, the implementation follows a simple bayesian inference model, using existing information derived from the SNP expression to evaluate the genes overall Allele specificity.

1.4 RNAseq based Transcriptomics

All projects presented here feature the use of RNA sequencing technology (RNAseq). In order to fully understand the impact of the technical possibilities and limitations, the mechanisms of sequencing need to be understood. Most sequencing approaches in modern transcriptomics rely on the machines and protocols developed by Illumina. Other approaches are available, but are currently less widely used. Ion-torrent for example produces RNA sequencing reads, but is less widely available, and was not used in any of the projects presented. PacBio is not commonly used for transcriptomics, due to its higher price, low throughput, and the relative small gain of information on the additional read length compared to the short but more cost efficiency of Illumina reads. Therefore, the introduction to these methods is omitted here. More modern methods like the upcoming oxford nanopore, or the Qiagen developed GeneReader as well as Thermo Fischers IonTorrent are not currently in wide use, and will also not be covered in detail.

Sanger sequencing and microarray-based transcriptomics

The basis of modern sequencing technology was developed by Sanger in two publications (Sanger and Coulson, 1975) (Sanger et al., 1977), and commercialized in the 1970s by Applied Biosystems. Based on the use of Polymerase Chain Reactions (PCR), it used a primer to target the elongation of a specific region. And like in any PCR, elongation is carried out via the introduction of nucleotides in vitro. To decipher the gene sequence, nucleotides with a fluorochrome labeling were introduced to a certain

ratio of the unlabeled nucleotides. Introduction of a labeled nucleotide terminated the elongation, and produced fragments of certain length with a terminal fluorochrome dideoxy terminator. Length measurement was carried out via capillary gel electrophoresis. Given different chromatic labels for the individual nucleotides G,A,T and C, the terminal nucleotides for any given position can be deciphered by measuring the intensity of the chromatic label at any given position, and positioned due to the length of the transcript. Sanger sequencing was the dominant sequencing technology until the beginning of the 21st century, and due to improvements that lead to automation, is still is in use today (commercially e.g as GATC sequencing). It was used for the qualitative expressed sequence tag (EST) discovery. Applications in quantitative transcriptomics were limited due to the very low coverage. sequencing was used to decode the gene sequence and combined with other methods for quantification. Microarrays were used to complement the known genomic sequences with an expression profile (Lockhart et al., 1996), (Bumgarner, 2013).

Microarrays were developed at the end of the 20th century. They use an affinity based labelling for nucleotide sequences and return a signal of intensity corresponding to the quantity of observed transcript. A problem with Microarrays is that the sequence analyzed has to be known beforehand, and most approaches rely on commercial versions containing e.g annotated genes in an organism (e.g (Malone and Oliver, 2011)). Gene independant transcriptomics was carried out via tiling arrays, analyzing the whole genome sequence in small overlapping windows. With microarrays, the intensity of expression is an almost arbitrary quantity. In theory, microarrays can be arbitrarily precise in their quantification, relative to a marker by extending the exposure time. Precision is limited by the detector, measuring the intensities of response and only up to a certain limit of biological binding site capability as well as cross hybridization. Comparisons of accuracy compared to RNA sequencing are not favourable (e.g (Zhao et al., 2014). Problems with microarrays were contrasted by the advance of sequencing technologies. The dependence of microarrays on known sequences is a limiting factor in exploratory analysis. This comes to bear especially if the investigation is based on more distantly related species, or

contains important single nucleotide polymorphisms. Additionally, since only intensity is measured, sequence diversity and expression rate were unified measurements without the ability to distinguish them. Any change to the sequence may reduce binding potential and therefore luminescence of the whole transcript. While RNA sequencing will give exact feedback on the changed nucleotides.

Next Generation Sequencing

An important distinction between classical genomic and transcriptomics sequencing is the experimental setup. Genomic sequencing was primarily used for de novo analysis of novel genomes. Transcriptomics is more closely related with re-sequencing approaches used later e.g to detect cancer driving singular mutations. In re-sequencing as with most transcriptomics, the reference sequence of the genome is known, reducing the need for long reads. This paved the way for the usage of short read sequencing, that were developed into transcriptomics sequencing based on sequencing complementary DNAs (cDNAs) of expressed transcripts. Illumina short read sequencing soon become the leading technology.

The trend towards quantitative sequencing began in the early 21st century. Three main approaches were developed to improve the original sanger sequencing. Reducing costs and improving speed, efficiency and output quality were the primary targets of the development. The first, in 2004 was pyrosequencing, developed as 454 by Roche. A demonstration summary can be found by Harrington and colleagues (Harrington et al., 2013). The commercialization by Roche was discontinued in 2013.

ABI SOLiD (Sequencing by Oligonucleotide Ligation and Detection) sequences became available commercially in 2006, introducing the concept of quantifiable sequencing. SOLiD sequencing was carried out by a bead clone strategy. Inducing many polymerisations of a single sequence on a bead increased concentration. The technology no longer relied on the dideoxy terminator approach, but introduced a ligase based detection assay. The application of SOLiD sequencing has regressed in recent years as the field has been taken over by Illumina. The first solexa sequencer was released

in 2006 and the company was bought by Illumina in 2007. This release has been followed up by rapid advancements in the technology and protocols. Originally reproducing genome sequencing, protocols for the first RNA sequencing were released in 2009. Currently, Illumina is the market leader in the transcriptomics field, with no real alternatives available. The illumina TrueSeq protocols, currently at TrueSeq 3 is the most used RNA sequencing protocol. This produces both a large standardization over the field, as well as a limit in methodological flexibility. The functional basis or RNA sequencing is the reverse translation of mRNA into cDNA. Technically, RNA sequencing with Illumina is carried out over the whole expressed regions of the genome. To limit the data, a common selection step is applied, removing all transcripts that do not carry a poly-A tail. Poly A tails are attached eukaryotic cells to processed mRNA (Sarkar, 1997), but not to ribosomal and other potentially noncoding RNA. For prokaryotes, where poly adenylation occurs less consistently, alternative rRNA depletion procedures do exist.

The Illumina/Solexa sequencers are characterized by: solid-phase amplification and a cyclic reversible termination (CRT) process, also termed sequencing-by-synthesis (SBS) technology. The sequencer can generate hundreds of millions of relatively short (36 — 100bp) read sequences per run. Modern sequencing libraries generate reads up to 250 bp in length, (2x 125 bp) in the high throughput HISeq (e.g TruSeq SBS V3), and up to 2x 300 bp in modern MISeq protocols (MiSeq Reagent Kit v3).

Common errors in NGS data

Common errors in NGS data are derived from the two aspects listed above, namely. Technical errors from the side of the machines, and experimental / biological errors from the library preparations or cultivations. RNA sequencing via Illumina sequencers is based on PCR amplification of transcripts. This introduces the sequence composition error. GC rich sequences will be transcribed to a lesser extend than GC poor regions. The same error occurs in Illumina genome sequencing, and is countered there, especially in de novo sequencing by introducing other sequencing technologies like PacBio, that contain a random error rate to fill the gaps

created by Illumina. The problem persists in transcriptomics, making gene comparisons less reliable by introducing the gene sequence as a source of variation in coverage. Problematic analysis are usually avoided by comparing identical genes. But earlier normalization approaches like Fragments (Reads) Per Kilobase of transcript per Million mapped reads (FPKM / RPKM) ignored this error source, to the detriment of their accuracy. Two main methods for normalization have been established, abundance and non-abundance based estimations. Notable non-abundance methods are TPM, RC, UQ, Med, TMM, DESeq, Q, RPKM, and ERPKM. Abundance based methods include Sailfish (Patro et al., 2013) and RSEM (Li and Dewey, 2011). A more comprehensive comparative analysis can be found by Li et al. (Li et al., 2015). Additional methods based on mass spectrometry derived linear correction used in this thesis include VST (Variance Stabilizing Transformation). Other technical errors are also well documented. For example 5 ends of read sequences carry lower quality values due to lower amplification and measurement accuracy. Errors in the library preparation kits are also documented. Combining advanced machines with complex chemistry provides error sources from the side of the machines. Several software suites are available to validate quality (FastQC), and remove the error as far as possible (e.g Trimmomatic (Bolger et al., 2014)). The current market, due to a lack of alternatives to Illumina does not allow for an accurate benchmarking. Another documented error concerns the Indel (Insertion/Deletion) sequencing mismatch in short read sequencing, misaligning sequencing due to a lack of reference accuracy. The combination of PCR and technical errors leads to high variability of coverage between adjacent loci, even if no transcription changes are expected. This variability ultimately has an impact on the overall coverage of genes, making the comparison of expressions of different genes less reliable than intragenic analysis such as differential expression.

1.5 RNAseq analysis

RNAseq Data analysis consists of a three step process: i) Preprocessing, which cleans the data, and ensures quality; ii) Processing, which essentially

turns the raw data of several giga-bytes into simple tables; and iii) Post-pocessing, which includes downstream analyses that produce meaningful results.

1.5.1 Preprocessing

Most projects described in this thesis concern the application of Next Generation Sequencing. More specifically, RNA sequencing. Data obtained from RNA sequencing is obtained as raw reads in fastq file format. This format contains the individual reads of a length determined by the library protocol, and the quality score for each base determined by the sequencing machine. The quality is given in a PHRED score, most commonly PHRED 33 format. In this format, ASCII characters, starting at ASCII character 33 (the symbol '!'), are used to encode the sequencers fluorescence detection into a quality score for the most likely base. The score translates to a probability of incorrect base call. A first step in handling this combined data is to filter the reads by their quality score. Quality scores in Illumina reads are not homogeneously distributed. A common observation is that read quality drastically decreases in the 5 direction of the reads. Quality is commonly assessed by the visualization software FastQC. After initial quality assessment, reads can be trimmed. Trimming here refers to the cutting off of low quality regions. In transcriptomics, the quantity of reads mapping to any region is important, little additional information is gained from longer reads. So theoretically reads over a length of 12 bp can be considered sufficiently specific. The reads are assessed by a sliding window approach, testing read quality in those windows, and removing areas below a certain quality threshold. The software solution chosen for the pipelines used in he work described here is trimmomatic v0.32 (Bolger et al., 2014), other options are available. Modern library preparations include the addition of specific adapters that can easily be removed from the reads, since their sequence is predefined by the sequencing protocol. This step is less critical for alignment based transcriptomics, but is generally recommended.

1.5.2 Processing

After successful trimming, the subsequent step in the quantitative analysis of expression is the association of each read to every single read to a region on the genome. Since all reads are cDNA derived from the mRNA inside cells, hypothetically all reads should have been derived from those cells. A usual step here, before the quantification of the reads, is a mapping step against the genomes of known potential contaminations. Especially when analyzing mammalian tissues, contaminant species like those of the genus *Mycoplasma* can be detected. Most work in my projects was carried out on yeasts, which themselves grow very rapidly, and do not commonly suffer from contaminations. Several different software solutions are available to map reads to their respective genome regions. The currently preferred tool is called STAR (Dobin et al., 2013, 2016). A few years ago, the most commonly used toolkit was the tuxedo suit, consisting of the short read mapper bowtie2, the splice junction mapper tophat2 (Kim et al., 2013) and several downstream analysis toolkits, cufflinks and cuffdiff (Trapnell et al., 2011), as well as cummerbund and others. For most projects described here, tophat2 and bowtie2 were used for mapping the reads to the genome. The mapping step associates each read to an area on the genome. After each read is mapped, quantification of areas of expression can follow. The core assumption in the analysis of RNA sequencing is that the amount of reads detected by the sequencer is proportional to the amount of mRNA that was produced by the cell. Additionally, the amount of mRNA is a result of the activity of the gene, and a general assumption is that expression correlates to the amount of protein for the given gene. The most common step is the counting of the reads overlapping each region of the genome. Several software solutions are available to estimate the coverage of annotated transcripts, depending on the complexity of the observed organism, different solutions are implemented. The one used most commonly, also in the thesis described here, is htseq-count. Htseq is a python package that deals with the quantification of read counts overlapping any region. The output of this analysis is the quantification of counts per transcript. An alternative approach to quantification is the estimation of all cellular transcripts. To accomplish this, the solution

implemented here follows the cufflinks transcriptome annotation approach. Given a reference genome, the transcriptome is aligned against the genome. Regions of sufficient expression with consistent coverage are returned as novel transcripts. This approach was used as a basis for the pipeline to detect long noncoding RNAs in *Candida parapsilosis* (see Chapter 2).

Postprocessing

Depending on the interest at hand, several approaches can follow. The most common direction is the quantification of expression and the analysis of what is called differentially expressed genes. Differential expression here refers to the detection of genes that have been actively, and more importantly, measurably changed in their level of expression. Due to the application of hypothesis testing, Differentially expressed transcripts are considered significantly active in expression. This activity is commonly related to the biological function of the transcript. Assessing differential expression is carried out via the comparison of gene expression over the whole transcriptome, the sum of all individual transcripts. The approach used most commonly in the described projects is the one implemented by DESeq2 (Love et al., 2014). The behavior of individual transcripts is modeled according to a negative binomial distribution. Variance of expression can be estimated by the analysis of expression over replicates. Knowing the variance for the individual genes, and the range of expression, a significance test for outliers against the expected distribution can be performed. This hypothesis test results in what is called a p-value. Using p- values has become a very common approach in biology.

The value tells for each transcript the likelihood that it does not behave differently from the null hypothesis. In other words, the likelihood of it being un-regulated (I.e not changing significantly the level of expression). P - values are probabilities given on a scale from 0 to 1, approaching but never reaching either. Due to convention, a probability of less than 5 %, or $p = 0.05$ is considered sufficient to make the claim, that a biologically meaningful occurrence was observed. Application of p-values in large scale analysis like transcriptomics leads to an important flaw in probabilistic.

Accepting 5 % of test results to be falsely returned as positive for the null hypothesis has little impact on individual measurements. But if all genes are assessed in a series of analysis, 5 % of false positives will result in large numbers of false positives. Yeasts of the saccharomycotina clade contain approximately 6000 genes, a 5% error results in approximately 300 false positive or active genes. To counter this trend, multiple testing corrections are carried out. DESeq2 (Love et al., 2014) uses approaches in multiplicity correction described above, which increases the p-value, based on the amount of repeated tests in the hypothesis scope. After the detection of actively regulated genes, the most common step involves the association of genes to their most likely function. This is most commonly done by associating genes to Gene Onthology (GO) terms. Several levels of certainty exist towards the gene association. In *Candida* species other than *C. albicans*, most genes are associated by sequence similarity to orthologs in either *C. albicans* or *S. cerevisiae*. Gene motif detections can be carried out to improve the accuracy of such computational predictions. Gene motifs contain conserved regions with known activities. Several motif databases exist for comparison. Yet the true function of those genes is not undisputed. Especially considering the evolutionary distance of those organisms, and their respective environments. Overall, GO terms can be used to estimate the biological meaning of an analysis, given sufficient numbers of associated genes are involved. In the most common approach, the active genes are tested for enrichment of certain functionalities against the background of all genes in the organism.

1.5.3 Developments since 2013

The advance of transcriptomics from the background of NGS is a recent development. Being introduced less than a decade ago, the analytical pipelines are not static, and massive changes have occurred even in the time span of this thesis. The pipeline described above was used from the early approaches in 2013 onwards, in order to preserve reproducibility. Yet the field has advanced and other methods should be mentioned. Most improvements have been implemented in the step of short read and splice junction mapping. Bowtie2 and Tophat2 are being replaced by more

advanced mappers, implemented in STAR (Dobin et al., 2016) aligner and HISAT2 (Kim et al., 2015). HISAT2 is based on the bowtie mapper and officially recommended by the Tophat2 developers as a more efficient tool. Differential expression analysis has not changed substantially in the last years. The same models are still being used. But other downstream analysis have to be considered more carefully. Some server solutions for GO enrichment analysis for example are being used even though they are outdated, leading to potentially misleading analysis (Wadi et al., 2016). Novel approaches are starting to appear, with the improved knowledge of cellular behavior, bayesian models, especially bayesian networks for transcriptomics data analysis are being implemented in approaches for transcriptome assembly and quantification (Kharchenko et al., 2014; Maretty et al., 2014)

1.6 Introduction to statistical approaches used in this thesis

Ultimately, the advances of measurement technologies lead to the introduction of the so called -omics techniques. An -omics technology refers to the analysis of a specific feature in the entire organism, or rather the collective analysis of an entire group of features. The core concepts of most work presented here are i) genomics; the study of all genes, ii) transcriptomics; the study of transcription and iii) proteomics; the study of cellular translation. Over the last years additional -omics techniques such as infectomics, interactomics, metabolomics and others have been introduced. An important point of the large scale observational concepts is the change in analysis compared to classical biology.

Mathematics and applied statistics are not fields that were widely used in classical biology. A famous article by E. O. Wilson written in his book *Letters to a Young Scientist* stated that [in biology] you dont need mathematics to do science. A concept that does not apply to modern -omics methods. Not only do modern quantitative methods require mathematical models to be used effectively, they often need large quantities of individual analyses. More recently Markowetz (Markowetz, 2017) argued that all biology has become computational biology. He argues that mathematical modeling, and

experimental design have long infiltrated biological approaches. Analyzing collectives of individual analyses is only possible with correctly applied statistics. Additionally, new problems arise if approaches from classical biology, like hypothesis testing are applied on larger scales.

In this section, a quick overview is provided concerning the approaches used in the subsequent Chapters. A special focus is given to statistical applications used in RNA sequencing data. Additionally, an introduction to the concept of bayesian analysis is described, referencing the project in Chapter 4.

1.6.1 Basic Concepts

Hypothesis testing

Hypothesis testing is the backbone of most comparative biological analysis. It is the primary tool in empirical research for detecting differences amongst groups of measurements. However, empirical research and therefore hypothesis testing have their limits. Since an empirical approach to research cannot eliminate uncertainty completely, it merely helps with the estimation of uncertainty (Banerjee Chitnis et al., 2011). The first step in hypothesis testing is the formulation of the competing hypothesis, followed by the estimation of correctness of empirical observation with data. Two types of errors derive from the empirical data under the expectations of hypothesis. Type I errors (alpha values) refer the the conclusion of a false positive, a false acceptance of the hypothesis. The claim commonly made is that group A differs from group B, and we are 95% certain that this is true. Type II errors (or Beta values) gives the complementary error for a false negative. The important consideration for biology is that hypothesis testing does not result in a clear yes or no answer, but merely the estimation of error for choosing one of the proposed hypothesis. In modern biology statistics has been standardized between experimental approaches, generating terminology that is followed by applied scientists without much consideration. Standard hypothesis tests return what is referred to as p- values, or the likelihood of committing a type I error. This applies most commonly to the rejection of the zero or null hypothesis, as the most common null hypothesis claims that

the two groups of measurements are identical. Rejection of this hypothesis is termed a significant difference. The commonly expected p-value in biology is 0.05, a 5% chance of false positives, in clinical experiments alpha may be set to 0.01 or below that e.g for drug discovery. The most common implementation is the t-test, a simple comparison of two normal distributions. Effects on multiple dependent groups is commonly tested via ANOVAs, ANalysis Of VAriances.

Multiple testing concepts and FDR

As described above, the most common implementation in biology accepts a hypothesis if the evidence against it allows for less than 5% uncertainty.In many biological experiments, multiple tests are carried out to analyze e.g individual genes for their significance. Three independent observations under the same 5% constraint have a combined probability of 14.2%.

$$(1 - (1 - \alpha)^n)$$

Meaning that with a 14% probability under the above constraints, at least one of the hypothesis was falsely accepted as true. Transcriptomics analyzes several thousand genes in a cell, with the expected rate of false positives around 5% of genes, potentially leading to a misleading analysis due to multiple testing. Several approaches have been proposed to counteract those false discoveries. The most common approaches deal with an adjustment to the p- value based on the amount of observations. Depending on the necessity of the analysis an approach that will sacrifice true positives for the sake of removing more false positives, or vice versa can be applied (Noble, 2010) (Aickin and Gensler, 1996). Complex statistical methodologies will have a general implementation of those approaches, and return an False Discovery Rate FDR (an alias to the Benjamini Hochberg correction) or p-adjust in DESeq2 (Love et al., 2014) instead of a simple p-value.

Correlation analysis

Correlation is a measure of similarity amongst samples behavior. The more identical two samples behavior according to an underlying expectation, the higher their correlation. Correlation analysis can be used to make predictions on underlying mechanisms by removing noise. There are two main approaches to compute correlation coefficients, the Pearson's product moment correlation coefficient, (Pearson coefficient) and the Spearman's rank correlation coefficient. A review on usage in a clinical setting was pubslished (Mukaka, 2012). Pearson correlations are commonly denoted as r for sample statistics. Pearson correlation coefficients assume a linear correlation between the samples. This is the method used for a project described in chapter 2 in associating gene function. In our approach we paired a pearson correlation function with a weighted clustering approach. The concept behind that analysis follows the idea that gene expression is co-regulated between genes of similar function.

Noise and how to deal with it

In order to obtain meaningful information from raw data in empirical research, it is not only important to find the right questions and hypothesis. It is also essential to understand effects that affect the measurement without providing information to the experiment. The term noise here refers to the unspecified errors that was detected by the measurements. In biological measurements, those errors derive from both the biological and the technical background. In complex analysis like RNA sequencing, noise affects individual bases, genes and the whole sequencing run in different ways. Several approaches have been proposed to deal with the noise in such analysis. The most common way of limiting noise, in transcriptomics, is normalization. By actively removing a known impact on the data variance it is possible to enable cleaner analysis pathways. The most commonly applied normalizations in RNA sequencing are Reads/Fragments Per Kilobase per Million Mapped Reads, RPKM / FPKM, proposed by Mortazavi, et al (Mortazavi et al., 2008). This normalization method has since fallen out of favour. The approach ignores some common high impact error sources,

and treats genes as a homogeneous block. As mentioned in a previous section, PCR based sequencing approaches have non random errors on their sequence, and do not normally show a homogeneous coverage. Current analytical pipelines suggest the more gene independent normalization of Transcripts Per Million TPM. A simpler normalization that ignores gene length. TPM is avoiding the length assumption from FPKM, and therefore the misleading assumption of sequence. A more complex normalization can be applied if more static underlying effects of several RNA sequencing runs have to be removed. E.g, as described in the chapters 2 and 4, samples were derived from different laboratories and combined into one analysis via a weighted clustering of expression. In order to remove effects derived from such data batches linear fitting algorithms are applied. In our case we opted for Variance stabilizing transformation as a methodology to stabilize variability of counts for genes on multiple sequencing batches. The method was described for RNA sequencing data by (Love et al., 2014). The original application derived from normalizing Mass spectrometry data, peaks, according to a linear graph to remove a baseline of noise. The original application of lowering a baseline to standardize output by keeping peaks stable was applicable to the RNA sequencing batch noise correction. The main benefit of this method is that it stabilizes the variance of batches of data for the downstream analysis without removing the individual genes expression ratios. Another way is followed by software suits for Differential Expression analysis, that use the noise to compute the expression variance expected for downstream classification.

1.6.2 Frequentist to Bayesian

Frequentist and Bayesian are two trains of thought in statistics. The one most commonly used in biology is Frequentist. It has the benefit of requiring simpler models which are easier to design and are consistent in their application. The simplicity of model comes at the cost of harder to interpret results. The most widely used of all biological applications of statistics is the Students t test. This test is based around defining observations into the shape of a normal distribution. A set of different observations, modeled in the same distribution can be tested to be different from the other

distribution. The test is called t-test, after the Student-t distribution and is applied very liberally on various biological processes. Results of such a hypothesis test is a p-value, already mentioned above, p values are not that easy to handle, and their meaning and implication are not grasped by the community at large. This leads to many misunderstandings in modern biology. Bayesian approaches, unlike frequentist rely on the existence of pre-existing knowledge, called priors. The underlying assumption is that any observation has to be evaluated in the context of its surrounding. Of course, this analysis can ever only be as good as the assumptions, or priors, available. It made therefore sense for modern biology to rely on frequentism, due to the limited assumptions required. As our knowledge of the systems we attempt to describe grows, bayesian analytics produce a more reliable analysis. In our approach described in Chapter 3, we implemented a bayesian approach for classification of Allele Specific Expression.

2 Long noncoding RNAs modulate virulence in the opportunistic yeast pathogen Candida parapsilosis

Long noncoding RNAs modulate virulence in the opportunistic yeast pathogen *Candida parapsilosis*

2.1 Abstract

The role of long, non-coding RNAs (lncRNAs) in fungi remains poorly understood. Most studies provide catalogs of predicted lncRNAs, which remain functionally uncharacterized. In this project, we combined transcriptomics and computational predictions to discover lncRNAs likely involved in virulence in the human opportunistic pathogen *Candida parapsilosis*. We used a combinatorial approach for the analysis of RNA-protein co-expression and predicted physical interactions to prioritize five lncRNAs for gene disruption and phenotipic characterization. All mutants exhibited various phenotypes, including mild to severe impairments of virulence. One of the selected potential lncRNAs was found to be sensitive to physiological temperature and oxidative stress, modulate genes involved in cell-wall permeability regulation and its deletion altered virulence in insect and mouse models. We identified the functional center that likely emerged from within an open reading frame, and which structure and function is preserved in the complementary strand. Our findings provide new insights into the origins and mechanisms of functional lncRNAs in yeasts.

2.2 Introduction

Candida parapsilosis, is an opportunistic human pathogen that ranks as the third most common source of systemic candidiasis worldwide, being responsible for roughly one third of neonatal *Candida* infections (Singh and Parija, 2012) (Pammi et al., 2013). This and other *Candida* species can be normal components of the human microbiome but, under specific circumstances, they can switch from commensal to pathogenic behavior, a process which involves changes in several cellular properties. Rapid shifts in morphology and physiology require complex networks of regulation, enabling individual cells to adapt to changing conditions. Under these

circumstances, regulation through RNAs may represent an advantage over protein-based systems, as RNAs can exert a function shortly after transcription. However, whether lncRNAs may play a role in human commensalism or virulence in *Candida* species is as yet unknown. The presence of lncRNAs has previously been reported in several other fungal species, most notably in the model yeast *Saccharomyces cerevisiae* (Yamashita et al., 2016). A recent study in this species attributes lncRNAs functions in mediating mating-type control of gametogenesis (Werven et al., 2013). Another study (Houseley et al., 2008) investigated the involvement of a non-coding, anti-sense transcript in the regulation of the GAL1-10 cluster. Only a few non-model fungal species have been investigated for their noncoding genome so far. These include the ascomycete *Trichophyton rubrum* (Liu et al., 2013) or the basidomycete pathogen *Cryptococcus neoformans*, where abundant lncRNAs have been predicted (Janbon et al., 2014). The presence of lncRNAs in several opportunistic *Candida* pathogens have also been examined. In a recent paper, Linde et al. (Linde et al., 2015) focused on the non-coding transcriptome potential of *Candida glabrata*, and Sellam et al. (Sellam et al., 2010) used tiling arrays and polymerase occupancy to produce an atlas of potentially pathogenicity associated noncoding RNAs (ncRNAs) in *C. albicans*. However, no downstream validation of the predicted functionality was performed. The original *C. parapsilosis* genome annotation did not include non-coding RNAs, although the possible presence of noncoding transcripts was mentioned in the last annotation update (Guida et al., 2011). More recently, Donovan et al. (Donovan et al., 2016) investigated ncRNAs in *C. parapsilosis*. However, this study focused on small nuclear RNAs (snoRNAs) and no downstream experiment tested the predicted functionalities. Thus lncRNAs in pathogenic fungi remain poorly investigated, with most studies being limited to large scale computational predictions based on transcriptomics data. Although some of them purport a possible role of lncRNAs in pathogenesis, we still lack concrete validated evidence. To accelerate the identification of functional lncRNAs potentially involved in pathogenesis we implemented a pipeline combining state-of-the art computational methods for lncRNA prediction from transcriptomic data and functional inference from RNA-protein coexpression and physical interaction networks. This, together with

information from a transcriptomics study of *C. parapsilosis* co-incubated with human THP1 monocites, provided a set of prioritized lncRNA candidates predicted to be involved in pathogenicity related processes. We constructed deletion mutants of five selected candidates, which were tested for an array of phenotypes. All tested candidates revealed altered phenotypes, ranging from impairment of virulence, sensitivity to copper and cadmium ions, to growth defects at 30°C. This underscores the validity and efficiency of our predictive and prioritizing approach, and suggests that our predicted catalog comprises several other putatively functional lncRNAs. One of the selected transcripts (MAD) was found to modulate temperature resistance and pathogenicity. We identified the functional center of MAD, which likely emerged from within an open reading frame (ORF) encoded in the opposite strand, and which secondary structure is conserved in forward and reverse orientation. Our findings provide new insights into the origins and mechanisms of functional lncRNAs in yeasts. To our knowledge this is the first study reporting experimental validation of lncRNAs related to virulence in a human fungal pathogen.

2.3 Results and Discussion

Predicted lncRNAs and their conservation, expression and structural characteristics

In a previous study we used RNAseq to monitor transcription of *C. parapsilosis* upon exposure to human undifferentiated THP1 monocytes (Tóth et al., 2017). This and other publicly available *C. parapsilosis* RNAseq datasets (a total of 106, See Supplementary Table S1), were used to produce a catalog of putative lncRNAs. For this, we used a state-of-the-art computational pipeline (see Online Methods) that i) predicts transcripts based on alignment of RNAseq reads to a genome reference and ii) infers the coding potential of predicted transcripts. To provide a base for comparison of our pipeline with other studies and benchmark it in a better studied organism, we applied our prediction pipeline to 166 RNAseq datasets

available for the model yeast *Saccharomyces cerevisiae* (See Supplementary Table S1). In total, 1097 $\pm$ 743 and 656 $\pm$ 406 lncRNAs were found in individual datasets of *S. cerevisiae* and *C. parapsilosis*, respectively (See Supplementary Table S1). This represents a low number, compared to the amount predicted by recent studies in multicellular eukaryotes e.g: (Iyer et al., 2015; Mallory and Shkumatava, 2016; Ariel et al., 2015). The relatively low amount can be justified given the relative lower number of physiological states, and the much denser genome of yeasts, compared to metazoans and plants. Predicted transcripts present across more than half of the available conditions were selected, resulting in 320 and 274 transcripts for *C. parapsilosis*, and *S. cerevisiae*, respectively. This subset was ranked according to predictive scores as well as occurrence over different replicates (see Online Methods), the upper quintile of this ranked list was selected for further study. This resulted in a subset of 64 lncRNAs in *C. parapsilosis* and 55 lncRNAs in *S. cerevisiae* (Supplementary Table S2). At the time of our study, 12 lncRNAs are annotated in *S. cerevisiae*, according to the Saccharomyces Genome Database (Cherry et al., 2012). Four of these (ICR1, PWR1 and RUF22, SCR1) were predicted in several of the *S. cerevisiae* datasets, of which SCR1 is found in the upper quintile subset as it was more consistently expressed across different conditions.

To assess whether our predictions were potentially derived from noisy transcription resulting from expression of nearby genes, we compared the patterns of expression of predicted lncRNAs to those of flanking protein coding genes. Our results (Supplementary table S3) show that the expression of predicted lncRNAs is not correlated with that of flanking protein coding genes, suggesting they are independently regulated. Indeed, the expression of our dataset of lncRNAs was less correlated to the expression of flanking genes than protein coding genes. Next, we investigated patterns in secondary structure, finding sequence stretches predicted to significantly more structured than a reshuffled sequence background (See Online Methods). Our results show that, similar to observations in humans (Yang and Zhang, 2015) yeast lncRNAs are less structured than coding mRNAs, but more so than intergenic regions. Finally, we assessed the level of sequence conservation by performing BLAST searches against the complete genomes of 14 different *Saccharomycotina* species (See Figure 2.1;

See Online Methods). We found that predicted fungal lncRNAs have overall low levels of sequence conservation, which are lower than those found in protein coding genes, but still higher than those of intergenic regions. The observed patterns of sequence conservation are comparable to those reported for lncRNAs in higher eukaryotes (Johnsson et al., 2014; Nitsche and Stadler, 2017), where some areas of the lncRNAs are well conserved, with interspersed nonconserved regions.

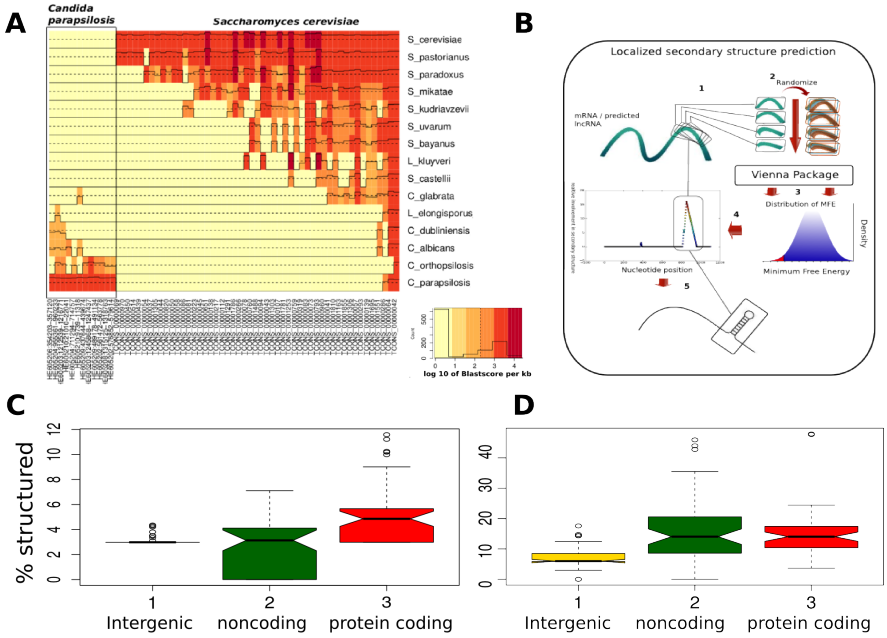

Figure 2.1: Figure showing the detailed analysis of lncRNAs. A.) shows the BLAST based analysis of sequence conservation, revealing a relatively low conservation score overall. B.) is a visualization of the pipeline used for localized secondary structure formation potential. It uses a sliding window approach to subset the sequence. Windows are tested against a randomized background seuqence and corrected for multiple testing. C & D show the relative amount of secondary structure compared to intergenic regions and protein coding genes.

Functional inference from networks of co-expression and predicted RNA-protein interactions

To assess potential functional roles of the predicted lncRNAs, we built transcriptional regulatory networks, including coding and noncoding transcripts, based on a weighted gene coexpression network approach ((Langfelder and Horvath, 2008), see Online Methods). The analysis resulted in two independent networks, one for *S. cerevisiae* and one for *C. parapsilosis*. We considered coexpression alone to be an insufficient criteria for the selection of lncRNAs for functional characterization in *C. parapsilosis*. The respective network was too dense to drive specific functional associations due to the limited amount of available distinct experimental conditions for the species (12). This resulted in a dense network with an average of 538 connections per lncRNA gene, as compared to three connections in the *S. cerevisiae network* (Supplementary File S2). Given the difficulty of gene deletions in *C. parapsilosis*, a diploid, asexual organism, we were interested in narrowing down our predictions to more specific ones. To do so, we filtered our predicted interactions with a layer of orthogonal information, by keeping only connections where physical RNA-protein interactions were predicted by CatRAPID ((Agostini et al., 2013), see Online Methods). This filtering step shifts the focus towards protein binding RNA, at the cost of limiting the prediction of other possible functionalities such as microRNA sponges (Liang et al., 2015), or lncRNA mRNA interactions (Jalali et al., 2013). Nevertheless, by focusing specifically on protein-RNA interactions that are both coexpressed and have high binding potential, we expect to remove a substantial amount of false positives from the overly dense networks. This step narrowed the focus to 17 lncRNAs from the original *C. parapsilosis* subset (Figure 2.2). Due to the status of *C. parapsilosis* as a nonmodel organism, its dense genome, and lack of evaluation for many possible genes, the presence of overlapping ORFs in the lncRNA sequences was considered, albeit it was not a criterion for exclusion. Using a functional enrichment approach we associated the transcripts to their most likely biological role (See Online Methods). The range of predicted functionalities included some features of special interest for the purpose of identifying pathogenicity related lncRNAs. Most notably adhesion or pseudohyphal growth. Five candidates were manually

selected for experimental validation, based on their predicted association with proteins involved in traits related to pathogenicity, and their expression in THP1 exposed cells. This included a lncRNA predicted to interact with genes enriched in transport (LNCRNA1), a lncRNA predicted to be involved in translation (LNCRNA2); a highly connected lncRNA with predicted regulatory and transport functions (LNCRNA3); one (LNCRNA4) predicted to interact with proteins whose orthologs in *C. albicans* regulate cell morphology and pathogenicity (IRS4 and SEC2)((Bishop et al., 2010; Badrane et al., 2005); and, finally, a transcript associated to transport (LNCRNA5).

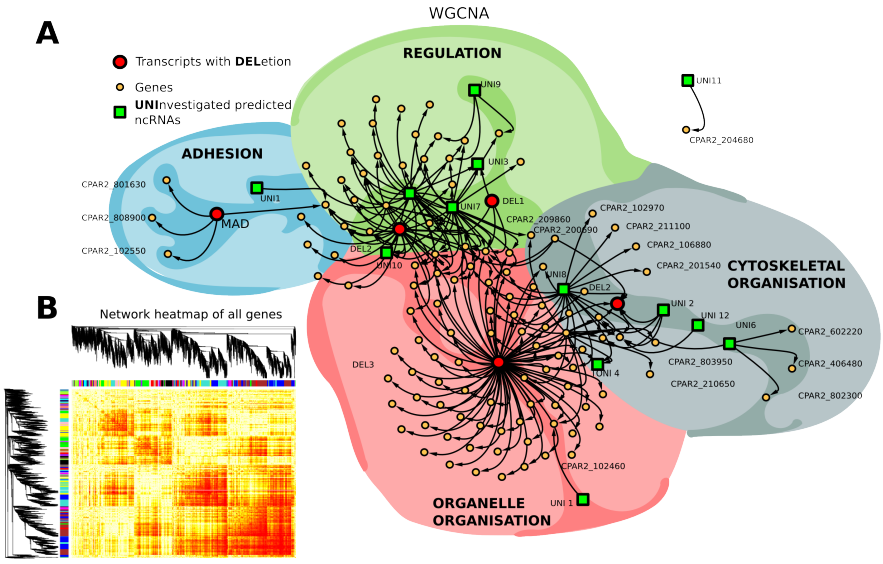

Figure 2.2: Weighted Gene Coexpression Network Analysis. A.) Shows the results of a WGCNA for *C. parapsilosis*. Background colors represent the enriched GO terms. The plot shows, in red circles, transcripts that were deleted and acreened. Additionally, uninvestigated transcripts are shown in green squares. Yellow represents expressed annotated genes. B) visualization of the WGCNA modules. Individual points represent expressed transcripts and their module association (color).

Deletion of selected lncRNA genes show diverse phenotypes

Following protocols established by Holland et al. (Holland et al., 2014), we created deletion mutants for the five lncRNAs described above (see Online

Methods). The resulting mutant strains and the wild type (wt) were then subjected to a battery of conditions, many of which are implicitly testing for the ability to survive and prosper in the human host (Figure 2.4). Notably, all five deletion mutants showed some degree of reduced virulence when tested in a *galleria* model of infection ((Németh et al., 2013), Figure 2.4). In addition, three of the mutants showed visible phenotypic changes under specific stress conditions, often in a temperature dependent manner (Figure 2.3). The lncrna3 Δ/Δ strain showed multiple phenotypes, including reduced growth in the presence of hydrogenperoxide and diverse cellwall stressors (congo red, caffeine, calcofluor white) at both tested temperatures (30°C and 37°C), with defects being more prominent at 37°C. Additional growth defects under various stresses (acidic pH, copper, iron depletion) were only observed at 37°C. Finally, lncrna3Δ/Δ strain showed reduced sensitivity to cadmium ions at both temperatures. The lncrna5Δ/Δ strain showed resistance to the cellwall stress induced by congo red in a concentration dependent manner, but only at 30°C. Finally, lncrna4Δ/Δ showed a range of interesting phenotypes: sensitivity to copper and cadmium ions, strongly reduced virulence in the *Galleria* infection model, and, importantly, the inability to grow on YPD agar at 37°C in the absence of additional stresses. Intriguingly the growth defect at 37°C disappeared when combined with osmotic stress, suggesting that response to this stress complements some putative functions of the deleted loci. Furthermore, this mutant showed increased sensitivity to several typical antifungal drugs, including the polyene drug Amphotericin B, as well as all tested echinocandins: Anidulafungin, Caspofungin and Micafungin. Interestingly, however, the sensitivity to one important antifungal drug, fluoconazole, was reduced in lncrna4Δ/Δ. This range of sensitivities of lncrna4Δ/Δ suggests major changes in cell wall properties and high sensitivity to various stresses. The inability to grow at 37°C degrees is expected to be relevant for pathogenic or even commensal behavior in the human host. We named this transcript MAD (Mdr1 AbD1 associated), and we will use this name hereafter.

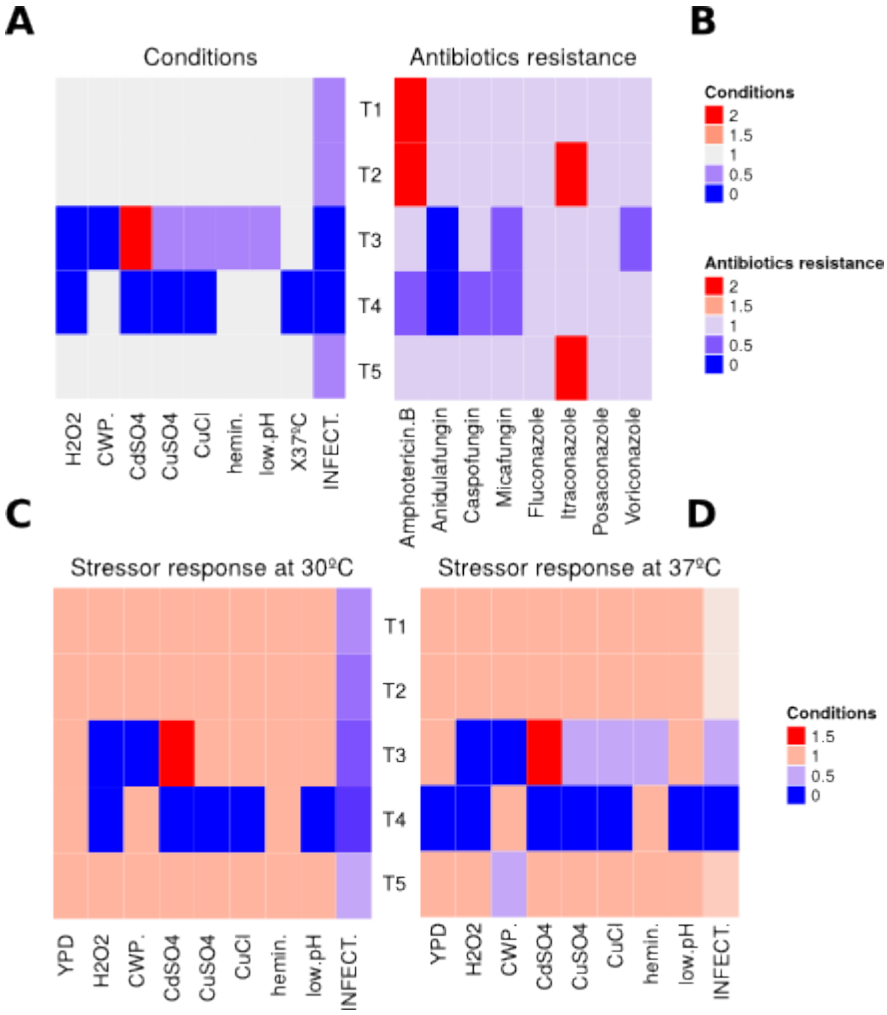

Figure 2.3: Figure showing the observed phenotypes on deletion mutants. In the plots A,C and D, colors correspond to relative growth to wild type. 1 refers to equal growth, lower values represent reduced growth. Figure B shows antibiotics resistance, Values correspond to relative MIC (Minimum Inhibitory Concentration) relative to wild type (1 = equal). A.) Combined phenotypes to various stressors. INFECT refers to *galleria* mortality, details in phenotype plot of Figure 2.4. B.) Antibiotics resistance in MIC relative to wild type. C & D) phenotypes observed at the different temperatures.

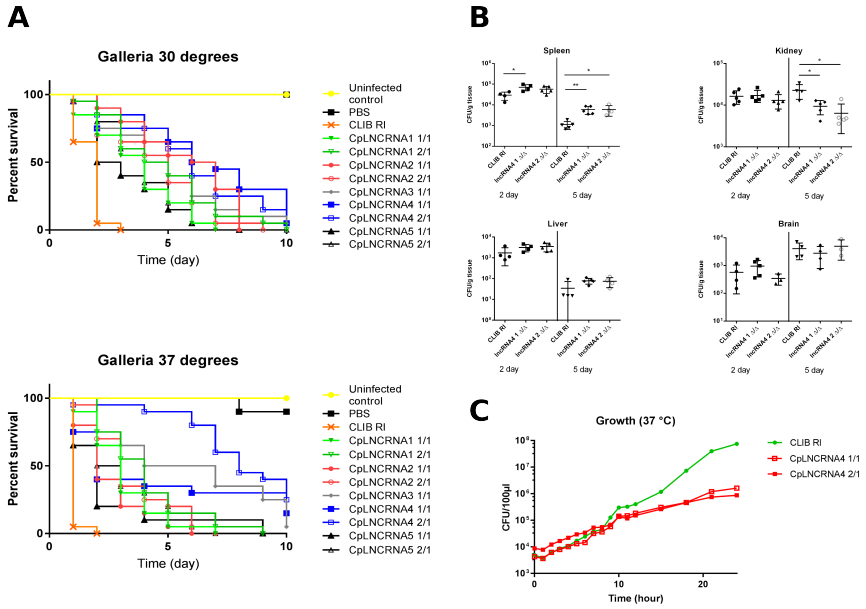

Figure 2.4: Pathogenicity Phenotypes. A.) *Galleria mellonella* infection study of all deletion mutants (4 mutants in dublicates + transcript 3, without replicate) compared to uninfected and PBS control. Lines represent the individual deletion mutants for the transcripts 1 to 5. CLIB RI represents to the wild type (mutant before deletion). B) Mouse results for the deletion mutant of transcript 4 (MAD) compared to the wild type (CLIB RI). CFUs (on log scale) were tested on day 2 and 5. Significance according to a t-statistic are shown in asterisks. C.) Growth curve of Transcript 4 (MAD) as CpLNCRNA1 deletion mutant against the wild type (CLIB RI). Growth in YPD medium at 37°C slows only after several hours.

MAD functional domain structure and function are strand-independent

Given the interesting phenotypes shown by the MAD knock out we chose this predicted lncRNA for further characterization. The analysis of sequence conservation and structure shows two areas of increased local secondary structure folding potential, one of which overlaps with a relatively highly conserved 200bp long region (Figure 2.5). A feature of initial concern was the presence of a putative open reading frame in the opposite strand where the lncRNA is encoded, which covers around 36 % of the MAD transcript and overlaps with the above mentioned conserved stretch. This

ORF is predicted to encode a transcription factor containing a basic Leucine Zipper domain, orthologous to *C. albicans* MET28. A comparison of synonymous and nonsynonymous mutations between *C. parapsilosis* and *C. orthopsilosis* MET28 loci suggests the protein coding capacity is constrained (see Online Methods). However, this comparison also reveals that a putative shared structure of the transcript is also constrained and maintained by compensatory mutations (p-value 0.01 based on structure conservation index from RNAz (Gruber et al., 2008). In *C. albicans*, MET28 has been found to be upregulated during mating (Zhao et al., 2005), while in *S. cerevisiae* it is known to mediate activation of sulfur metabolism genes (Kuras et al., 1996, 1997)). These potential functions, and the range of phenotypes for the null mutant described in *S. cerevisiae* are not related with those observed here for lncrna4Δ/Δ (hereafter MADΔ/Δ). The RNAseq datasets used to predict MAD are not strand specific, precluding us to differentiate expression from each of the two DNA strands. However, expression analysis over the range of conditions available shows no significant changes in the expression levels of the ORF region compared to the whole transcript. That is, in the available conditions the region covering the ORF seems to be not expressed independently from the rest of the region covering the MAD transcript. We tested strandedness in CLIB WT exposed to THP-1 phagocytes by using qPCR (See Online Methods), which showed roughly ten-fold ($9.83 \pm 2.3 : 1$) higher expression of the MAD bearing strand as compared to the ORF strand (see Supplementary Table S1). An investigation into potential functional centers overlapping the ORF is documented in the Online methods. We nevertheless designed an experiment to specifically test whether the absence of the lncRNA or the encoded protein were responsible for the mutant phenotypes (37°C). For this we tested the effect of reintroducing, in the knock out background, i) the native ORF sequence under a promoter in forward direction, ii) the same sequence under a promoter in reverse direction (i.e. the strand of the MAD transcript) and, iii) a modified ORF with a frameshift mutation in forward direction (see Online Methods). All three introductions rescued the temperature sensitive phenotype, indicating that transcripts from the forward, frameshift and even the reverse strand of this region are able to complement the function but that, importantly, the functional complementation seems to be largely

independent of the ORF translation (Figure 2.6). Closer analyses revealed that identical structures can be formed in forward or reverse strands by the genomic region comprising the previously described functional center (see Figure 2.5). To our knowledge, this ability to fulfill the same function by transcripts encoded by complementary strands has not been reported for a functional lncRNA. As we will discuss below this observation may be related to the evolutionary origin of MAD. However, in the above experiments we could not fully exclude the possibility that an alternative start codon produced a truncated, functional protein. We thus integrated another construct which included a stop codon after the last possible methionine residue (see Online methods, Figure 2.6(B)). This construct was not able to restore growth at 37°C in the MADΔ/Δ background, although the finding that this construct expressed the transcript only at 30% of the wt expression of MAD, made this result also not conclusive. We thus must remain cautious about a possible dual role of this locus in the observed phenotype.

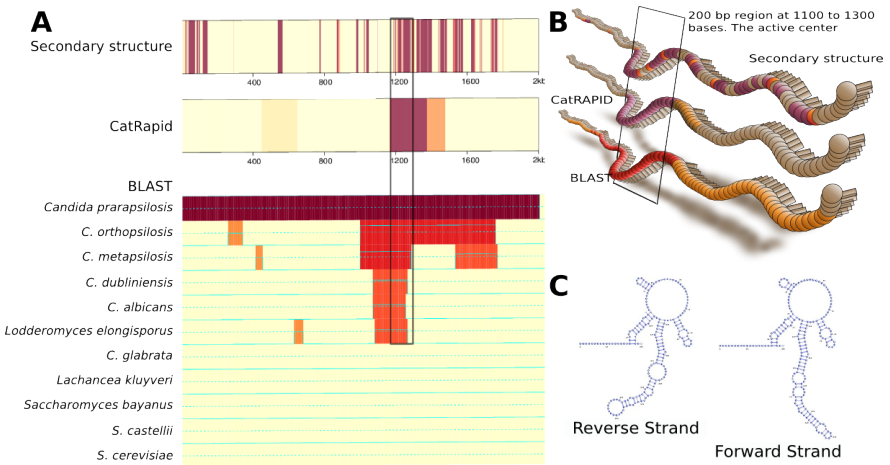

Figure 2.5: Functional center evaluation. A.) The figure shows the triplot of overlapping significant regions for MAD. The sequence is displayed from left to right, covering 1986 nucleotides. The Secondary structure was analyzed according to the pipeline described above. BLAST results were analyzed according to the approach discusses in the methods. B.) Visualization of the three approaches, with a potential inferrence of the functional center. C.) Predicted secondary structures for the forward and reverse strand for the region highlighted in B according to Beagle web server

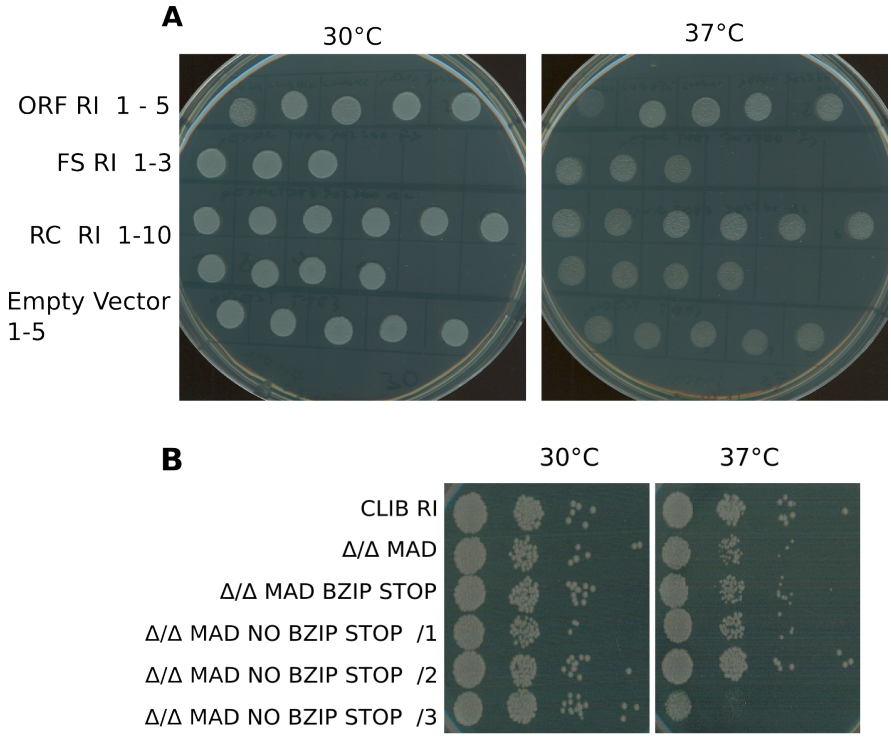

Figure 2.6: Reintegration results. A.) Shows the various results of ORF, Frame Shift and Reverse Complement reintegration. All transcripts complement the phenotype compared to the empty vector (bottom row). B.) Shows the results of the second stop codon reintegration. 30°C shows normal growth, while at 37°C deletion mutants show at least a partial rescue compared to the empty vector.

Deletion of MAD alters expression patterns upon growth at 37 degrees

To gain further insights into the potential role of MAD in physiological temperature tolerance, we performed strand-specific RNAseq experiments with the MAD mutant and wt strains grown at 30 and 37°C (see Online Methods). Our results show that both strains grow similarly during the first 8h of exposure to 37°C, after which a retention of growth in the MAD knock out sets in (supplementary table 3). To obtain homogeneous cultures, we took samples after 15 hours of growth at 30 and 37 degrees for RNA extraction (see Online methods). Reassuringly, in the wt strain the expression of the strand coding for the ORF was negligible at both

temperatures, and roughly three orders of magnitude smaller than the MAD strand, further supporting a functional role for the lncRNA in this condition. We first compared, in each of the genetic backgrounds, genes significantly up- or -down regulated at 37°C degrees as compared to growth at 30°C. This has the potential to inform how the two strains react to different temperatures. In the wt strain 77 genes altered their expression between the two temperatures (basecount $>$ 100 & log2fold of $<$ |1.5 |). This number was severely reduced to 47, of which 11 are unique to the deletion mutant MADΔ/Δ. This suggests an impaired regulatory activity. The list of differentially regulated genes (see Supplementary list 2) reveals the presence of several genes related to mitosis (APC1), ion binding (SUT1), transcription regulation (SSN8) or external membrane processes and transmembrane transport (YOR1, MDR1). Of note, during exposure to 37°C MDR1 is downregulated in wt, a plasmamembrane transporter responsible for effective removal of fluconazole from the cytoplasm in *C.albicans* (Lamping et al., 2007). This regulation was lost in the MAD knock out, which is in line with the observation of reduced sensitivity to fluoconazole as compared to the wt strain. Although there was no difference found in the expression of MDR1 at 30°C between the two strains, our observation suggest a MAD dependent MDR1 regulation that might take place upon fluconazole exposure. Comparing the wt and MADΔ/Δ gowing at 30°C on YPD, only 7 genes are differentially regulated. The most interesting one is CPAR2_808130 (ABD1 in *C. albicans* and *S. cerevisiae*), a gene with orthologs involved mRNA maturation, mRNA capping and a cellular response to drug that co-localizes with the DNA-directed RNA polymerase II holoenzyme (Pol II). ABD1 is an essential protein in *S.cerevisiae*, (Schroeder et al., 2004), and considered gene-specific RNA pol II transcription factor (Schroeder et al., 2004), especially since other pol II activating genes were differentially expressed in the above comparison (SSN8). We next compared, for each tested temperature, the genes that had significantly different expression in each of the backgrounds. This directly measured the effect in overall transcription of the MAD deletion. Considering the above cutoff criteria (basecount $>$ 100 & log2fold of $<$ |1.5 |) there was only one gene, CPAR2_106960 showing altered expression that was downregulated at 37°C in MADΔ/Δ compared to wt. This suggests that the presence or absence of

MAD has a very minor impact on cell growth at 30 degrees. In contrast, growth at 37 degrees shows 11 differentially expressed genes. Among them there are a few well studied ORFs, including MDR1 the Multi Drug Resistance gene 1 (Lamping et al., 2007), the product of SLS1 is involved in aerobic respiration and CPAR2_807070, the ortholog of *C. albicans* SSN8 that is a conserved component of the eukaryotic transcriptional machinery and provides a connection between regulatory elements and the RNA polymerase II. ((Bryan et al., 2002), (Lindsay et al., 2014), (Boube et al., 2002).

MAD deletion alters virulence patterns in a mouse model

To assess the potential significance of the observed defects in virulence towards a mammalian host, we set out to test the effects of the MAD mutant on virulence in a murine model of candidiasis (see Online methods). We measured colony forming units (cfu) after two and five days of bloodstream infection, in liver, spleen, kidney and brain tissue (Figure 2.4). We observed that temperature sensitivity is not lethal in vivo and that, similarly to the effect of osmotic stress, physiological conditions in some tissues may even allow growth. As compared to the wt strain, MADΔ/Δ infections resulted in similar loads in the brain and liver, and, surprisingly, a higher load in spleen. In contrast, a severe depletion was observed in kidney, which is the organ most affected in *Candida* bloodstream infections in this murine model (Hebecker et al., 2016). Altogether, these results indicate a function of MAD in pathways that modulate virulence and survival in the host.

2.4 Conclusion

We implemented an efficient workflow to accelerate the discovery of lncRNAs implicated in pathogenesis, as illustrated by the finding that all selected potential lncRNAs genes produce a phenotype when deleted. comparable study of deletion mutants by Holland et al. found only 37% of deletion mutants show phenotypes affecting virulence directly or indirectly

(Holland et al., 2014). Therefore, the co-expression based pipeline shows potential for the analysis for other cellular features, such as genes. Based on the interesting phenotypes shown, we have focused on one of the tested candidates (MAD). Deletion of MAD results in sensitivity to copper and cadmium ions, strongly reduced virulence in the *Galleria* infection model, and the inability to grow on YPD agar at 37°C. Although the presence of an overlapping ORF in the opposite strand of MAD was of initial concern, our experiments suggest that the phenotypic complementation obtained by reintroducing this genomic region is likely independent of the translation of the ORF. In addition, the finding that the most conserved and structured region of MAD is predicted to form similar structures when transcribed from the two complementary strands, and the fact that both transcripts are able to complement the phenotype of the knock out in a manner independent of the translation may not be unrelated. We hypothesize a plausible evolutionary scenario for the origin of MAD, in which this lncRNA originated from co-option of a protein-coding transcript. In such scenario a regulatory function through RNA-protein interactions may have emerged as a secondary role of a transcript of a functional MET28 gene in *C. parapsilosis*. Provided the resulting transcript could be folded in a similar secondary structure, anti-sense expression of the MET28 locus would have ensured the presence of the regulatory function, independently of the transcription and translation of MET28, thus avoiding potential deleterious effects of the expression of this transcription factor. Further extension of the anti-sense transcript may have served to optimize the alternative function. We hypothesize that such a scenario may have also played a role in the origin of other lncRNAs in other eukaryotic species, implicating that many lncRNAs start their lives as moonlight functions of bonafide mRNAs. This would explain observations of abundant anti-sense lncRNAs that overlap with protein-coding genes. An alternative potential explanation involves a role of MAD in regulation of chromatin modeling. It has been shown in *drosophila* (Wilusz et al., 2009) that noncoding RNAs like *Kcnq1* (Smilinich et al., 1999) can play a role in chromatin structure. In this case, transcription of the antisense transcript may be useful for generating access to dense chromatin regions. Admittedly, we cannot fully exclude a possible role of the ORF sitting in the opposite strand in the observed phenotypes. Given that the predicted

functional center of MAD overlaps with the ORF, and the structure can be formed in RNA transcribed in either sense, it is challenging to disentangle the contribution of each molecule. However, considering the much higher expression of the MAD strand in the analyzed conditions, and the fact that the knock out phenotype is complemented by constructs with a frameshift mutations, and under the control of promotors driving the expression in the sense of the lncRNA, makes us support an important role for the lncRNA, albeit perhaps not exclusively.

2.5 Online Methods

Data

We downloaded 166 RNAseq data sets from 13 different experiments *S. cerevisiae* and 48 RNAseq runs from 1 experiment (Guida et al., 2011) for *C. parapsilosis* from SRA (Leinonen et al., 2011), encompassing 19 different conditions see Supplementary Table 1 for details. Supplemented with 58 RNAseq runs from 3 experiments produced by our group for *C.parapsilosis*. Reference genomes and annotation files for *S. cerevisiae*, *S. paradoxus*, *C. parapsilosis*, *C. metapsilosis*, *C. orthopsilosis*, *C. glabrata* and *C. albicans* were obtained from SGD Saccharomyces cerevisiae, *S.pastorianus*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. uvarum*, *S. bayanus*, *S. castellii*, *Lachancea kluyveri*, from SGD (Dwight et al., 2004) and *C. glabrata*, *Lodderomyces elongisporus*, *C. albicans*, *C. dubliniensis*, *C. parapsilosis* and *C. orthopsilosis*. from CGD (Arnaud et al., 2005).

Transcriptome Assembly

For initial data preprocessing we used Fastqc 0.10.1 (Andrews S., 2010) for quality assessment and Trimmomatic v0.32 (Bolger et al., 2014) conditions: LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 for discarding low quality regions and reads. Samples with more than 15% of discarded

reads were dismissed from the analysis. The Tuxedo suit was used on Illumina datasets exclusively. Strain specific genome reference sequences were available, and were used for a Reference assisted assembly, in None of the cases, de novo assemblies were necessary. For the reference-assisted assemblies, individual samples were assembled with tophat 2.01.13 (Kim et al., 2013) with default settings, and mapped with the bowtie 2.2.4 (Langmead and Salzberg, 2013) short read mapper Cufflinks 2.2.1 (Trapnell et al., 2011) against the respective reference sequence. Subsequently, cuffmerge was used to merge the individual samples for transcriptome prediction. Using mostly paired end data, replicates were merged, using cuffmerge, to contain a total of about 20 to 30 million reads. Increasing the read amount further did not increase the quality of the assembled transcriptome.

Prediction of lncRNAs

LncRNA prediction was done computationally on the assembled Transcriptomes. In a first step, Coding Potential Calculator CPC, (Kong et al., 2007) with default parameters was used to predict lncRNAs de novo. CPC uses a support vector machine aided blastx approach against an existing protein database, in our case Uniprot Ref90 (UniProtConsortium, 2015), to select transcripts that are not similar to existing proteins, and show limited open reading frames. This approach provided reproducible predictions across the available transcriptome assemblies. In a second step, Coding Potential Assessment Tool CPAT (Wang et al., 2013) with default parameters was used to predict lncRNAs from codon usage frequencies. Based on a learned regression model, this program provides a linguistics based alternative to CPC, with vastly increased speed. CPAT can accurately predict coding potential on the patterns of usage of hexameric sequence occurrence. To predict sequences, a training set of coding and of noncoding RNAs has to be provided. Due to a lack of noncoding prior information, sequences from intergenic regions situated more than 1000 nucleotides away from genes, were used as priors for non-coding. The most notable difference between the two methods lays in the existing requirements, and the computational expense. CPAT requires more prior information in the form of existing non coding re-

gions. While CPC is very resource intensive due to the use of massive BLAST searches. The outcome of both prediction tools overlapped significantly. 92.3 $\pm$ 5.7 % of CPC predictions could be validated with CPAT predictions. Only sequences predicted individually by both approaches were kept as possible targets.

Differential expression

After mapping with Tophat2 (Kim et al., 2013), DESeq2 (Love et al., 2014) was used to calculate differential expression of the whole transcriptome. After calculating the log2 transformed changes against reference base conditions, we used Pearson correlation coefficients to compare expression patterns over the different conditions. To enable a better comparison between different experiments in regard of technical error sources, a normalisation via variance stabilizing transformation, implemented in the DESeq2 (Love et al., 2014; Anders et al., 2010) package was carried out for downstream analyses.

Sequence conservation between closely related species

We performed a similar approach as the one described by Ulitsky (Ulitsky et al., 2012). This approach basically uses Blast comparisons to investigate sequence conservation of the predicted subsets of lncRNAs over several species within the saccharomycotina clade, namely *Saccharomyces cerevisiae*, *S. pastorianus*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. uvarum*, *S. bayanus*, *S. castellii*, *Lachancea kluyveri*, *C. glabrata*, *Lodderomyces elongisporus*, *C. albicans*, *C. dubliniensis*, *C. parapsilosis* and *C. orthopsilosis*. Our analysis focused on the sequence comparison of the whole lncRNAs as well as fragmented lncRNAs of 100bp against the available transcriptomes. An e value of 10e-5 was considered significant. For comparison, we tested sets of randomly selected intergenic regions and annotated protein coding genes, normalized over the length of the region to blast score per kilobase.

Secondary structure analysis

We predicted the propensity to form secondary structure using RNAfold 2.1.9 (Gruber et al., 2008). In a first attempt, the secondary structure of whole lncRNAs was tested. To obtain a background model, the individual sequences were randomized and individual predictions were carried out on 50 different randomizations. The 50 randomized predictions were used to define a randomized background population of structures, and an upper tail test of population mean was carried out, applying a cutoff of p 0.05 for significance. Since lncRNAs function is sometimes postulated to be mediated by local secondary structure formation, such as specific binding sites, we developed a simple pipeline, testing for secondary structures in sequence sliding windows of up to 50bp in length. The pipeline, described in the Online Methods, tests the increased likelihood of secondary structure formation of a given sequence over a randomized one. We applied a variation of a pipeline designed by Yang and colleagues (Yang and Zhang, 2015). Yang et al. Used a combination of window prediciton via RNAfold (Zhao et al., 2005) and PARS experiments to compute the sequence wise relative structure folding potential, in terms of an abstract score. Yet even by using PARS data, were unable to give a per nucleotide binding probability. No PARS data is yet available on any of our predicted regions, so in our case as well, no additional value could be gained on a per nucleotide prediction. Instead the re-translation from windows to nucleotides would introduce additional noise into our prediction. We decided instead to work with the more abstract whole-window predictions, and comparing them directly to a randomized background model. Choosing a window size of 50 nucleotide masks all but the shortest secondary structures. Yet it increases reliability of the prediction, and shifting the 50 nucleotide window by each nucleotide results in a 1 nucleotide resolution of relative probability of secondary structure involvement. The same method, an upper tail test of population mean against 50 nucleotide windows of 50 randomized sequences of the same lncRNA was applied. A cutoff of an (adjusted via Bonferroni Holm) $p < 0.05$ was considered significant. To correct against multiplicity, the Bonferroni-holm method was applied (Aickin and Gensler, 1996). The pipeline was run against our predicted long noncoding RNAs, as well as 50

annotated protein coding regions, and randomly selected intergenic regions, longer than 200 bp and more than 200 bp from upstream and downstream protein coding genes.

Functional assessment

We used weighted gene correlation network analysis (WGCNA), as implemented in the R package WGCNA (Langfelder and Horvath, 2008), to create a network of correlating gene expressions from our available RNAseq data. To correct against inter-sample variance, variance stabilizing transformation, described in the DESEq package (Anders et al., 2010) was applied to the raw counts of our data over all available conditions. WGCNA requires the input of a chosen soft-threshold power. For our experimental data, a Power of 12 provided an acceptable clustering into 10 modules. After the creation of the initial networks, in the case of *C. parapsilosis*, additional masking had to be applied, to remove incidental network connections resulting from a small sample size. For this, we used the catRAPID software (Agostini et al., 2013). CatRAPID omics calculates the interaction propensities of a RNA against the proteome. It returns Discriminative Power (DP) ranging from 0% (unpredictability) to 100% (predictability), the recommended threshold to assume interaction is 75%. Since our work was focused on non-model organisms, we increased the threshold for our samples to significantly higher levels of 90%. The resulting RNA-protein interaction predictions were used as a mask for the WGCNA, removing all interactions that were not predicted independently by both methods. This resulted in a much smaller subset of interactions. Due to the increased sample size on *S. cerevisiae* this step became unnecessary. Network analysis and display were performed using Cytoscape 3.1.1 (Shannon et al., 2015). The obtained networks were analyzed for GO enrichment of both functional and localization within the cell using the Candida Genome Database toolset [CGD (Arnaud et al., 2005)]. For *Candida parapsilosis*, the functional information was used to select interesting targets for a knock out study.

Functional center investigation

To evaluate the potential functional center of MAD, we used a BLAST based approach to investigate its conservation, as shown by Ulitzky et al. (Ulitsky et al., 2012). Conservation was estimated against the reference genomes of *saccharomycotina* listed in the Online methods. We found an area in the transcript, overlapping the Open Reading Frame to be significantly conserved (Blast cutoff e-5). Results are shown in Figure 2.5. Subsequently, we investigated the secondary structure formation potential, using the pipeline described above (details in the Online Methods). Several windows of increased secondary structure were found, displayed in Figure [Functional Center]. Using the webserver WebBeagle (Mattei et al., 2015), we compared the secondary structure by windows of 200 nucleotides. The window of 1100 to 1300 bases showed a significant similarity between its forward and reverse strand (Z value 3.74). Lastly, we compared the overlap with the predictions for binding potential by CatRapid. (Agostini et al., 2013). Two binding sites were identified by CatRapid, shown in Figure [Functional Center]. We conclude, that due to the conservation score, the local secondary structure folding potential and the predicted protein binding sites, the functional center of MAD1 is most likely located between base 1100 and 1300.

ORF Expression evaluation

MAD overlaps an open reading frame, that shows homology to MET28 in *C. albicans*. To evaluate the importance of the ORF, we first investigated the expression of the sense (ORF) and antisense (MAD). Transcription analysis via qPCR and strand specific sequencing showed respectively that a ratio of 90.4 $\pm$ 1.7 % (qPCR a ratio of (9.81 $\pm$ 2.3) : 1) and 99.6 $\pm$ 0.11 % (RNASeq inferred from count data) of transcription occurred on the antisense strand. We consider RNAseq to be the more quantitatively accurate method. Although the expression of the Antisense strand actively increases at exposure of 30°C (Δ expression around 26%), the sense strands

expression remains below the detection limit, with less than 10 reads mapped against the full ORF.

Reintegration results

To investigate the importance of the ORF, we introduced three variations of the fragments back into the deletion mutant. We reintegrated the transcript in individually in forward (MAD), reverse (ORF) and with a frameshift in the initial start codon into the NEUT5 locus [see Online Methods]. Results are shown in figure 2.6. The results are not clear cut. Our analysis show that both the forward and reverse strand of the ORF complement the function. This is in line with our observation that transcripts of both strands form a significantly similar secondary structure. Reintegration of the ORF with a frame shift mutation in the initial methionin also complemented the phenotype. An important consideration is whether the frameshift introduced into the first Methionin leaves a further potential starting point to be translated downstream in the transcript. This would potentially leave the transcript to be translated from a subsequent methionin and produce a functional transcript. To analyze this, we introduced a nonsense mutation into a position within the transcript. Details in figure 2.6). This phenotype did not complement the function, it reduced growth under 30°C compared to the deletion mutant. The analysis of the ORF expression shows that the new integration is only expressed to 30% of the other complementing transcripts. A re-integrand of the full ORF in equal conditions complemented the phenotype in some cases.

Growth conditions

Strains for transformation and phenotypic characterization were grown in YPD (0.5% (m/V) yeast extract, 1% (m/V) peptone, 1% (m/V) glucose) liquid media supplemented with 100 unit/ml penicillin-streptomycin antibiotics at 30°C with vigorous shaking. For phenotypic screening cultures were synchronized. Hetero- and homozygous deletion mutants were selected on

selective dropout solid media containing 0.19% (m/V) yeast nitrogen base, 2% (m/V) glucose, 2% (m/V) agar with 100 unit/ml penicillin-streptomycin, supplemented with L-Leucine to a final concentration of 0.5 mg/ml for heterozygotes. For dominant selection, nourseothricin was applied at a final concentration of 100μg/ml in YPD media (described above) supplemented with 2% (m/V) agarose.

RNA extraction

RNA was extracted by using Thermofisher RiboPureTM RNA purification kit according to the manufacturers instructions. When RNA was extracted from yeasts exposed to phagocytes, the cells were collected, suspended in ice-cold nuclease free distilled water and forced through a 29G syringe five times. The homogenate was washed with ice-cold nuclease free distilled water and the yeast pellet was used for RNA extraction. RNA integrity was checked with AgilentTM 2200 TapeStationTM Instrument.

Mutant generation

Deletion mutants were created with the double auxotrophy complementation combined with fusion PCR method adopted for C.parapsilosis by Holland et al. (Holland et al., 2014).Using first the HIS1 then the LEU2 cassette. Two independent homozygous mutants were generated for both loci (except CpLNCRNA3). Heterozyogus mutants were checked only by colony PCR, homozygous mutants were verified with PCR and Southern blot by using DIG labeled probes specific for HIS1 and LEU2 markers (detailed in (Gácser et al., 2007; Holland et al., 2014)). Genomic DNA was digested with BamHI ($\Delta\Delta$lncrna1-4) and PvuII ($\Delta\Delta$lncrna5).Reintegrant mutants were generated by using InvitrogenTM GatewayTM system. pDEST-TDH3-URA3-RPS1-GTW was a kind gift from Professor Christophe dEnfert (Chauvel et al., 2012). URA3 selection marker was replaced to NAT (nourseothricin acetyltransferase) that was amplified from (PLASMID) with restriction sites SpeI and SacI. pDEST-TDH3-URA3-RPS1-GTW was digested with SpeI and SacI

to remove URA3, purified and ligated with SpeI and SacI digested NAT marker to create pDEST-TDH3-NAT-RPS1-GTW. An intergenic region, CpNEUT5L (named after the ortholog of *C. albicans* NEUT5L) was chosen to target (Gerami-Nejad et al., 2013). A 758 bp region was amplified in two fragments to artificially introduce a StuI site in the middle with fusion PCR similar to that of Chauvel et al. introduced in the RPS1 region (Gerami-Nejad et al., 2013). The fragment was supplemented with flanking MluI and SpeI sites. The pDEST-TDH3-NAT-RPS1-GTW and the CpNEUT5L fusion product were digested with MluI and SpeI, gelpurified and ligated together to create pDEST-TDH3-NAT-CpNEUT5L-GTW. Fragments for BP cloning were amplified from CLIB WT gDNA by using Thermo ScientificTM PhusionTM Hot Start II High-Fidelity DNA polymerase, and were purified with PEG8000-MgCl2 method. GatewayTM cloning process was performed according to the manufacturers instructions. For BP cloning pDONR221, for LR cloning pDEST-TDH3-NAT-CpNEUT5L-GTW plasmids were used. pDONR and pDEST plasmids were propagated in *Escherichia coli* DB3.1, pENTRY and pEXPRESSION vectors were transformed into *E. coli* 2T1. Reintegrant mutants were screened by colony PCR (CPAR2_702260_FOR_ReTi and Colony_check_REV), and Southern-blot, see above. Genomic DNA was digested, with EcoRI, CpNEUT5L Downdown DIG labeled probe was used for hybridisation.

Antibiotics resistance

We tested minimum inhibitory concentrations for three common types of antifungals: polyenes (Amphotericin B), echinocandins (Anidulafungin, Caspofungin, Micafungin) and azoles (Fluconazole, Itraconazole, Posaconazole, Voriconazole). Experiments were performed in 96 well plates in three parallels. Antibiotics were diluted in a two fold serial dilutions with RPMI-MOPS media in a final volume of 100 μl. 100 μl yeast suspensions in RPMI-MOPS containing 2000 cells were added to the wells, and incubated at 30°C for 24 and 48 hours. Plates were observed, and the lowest antibiotic concentrations with no noticable yeast growth (MIC) were documented.

Phenotypic analysis

Synchronised cultures were washed twice (2400 xg, 3 minutes) and resuspended in 1x sterile PBS then counted with Burker-chamber. Dilution series were prepared at 104, 103, 102, 101 cells/5μl concentrations. Suspensions were pinned onto agar plates representing different stress circumstances or growth conditions (Supplementary). Plates were incubated at both 30 and 37 °C degrees (YPD only plates were at 20, 25 and 40) and scanned after 2 days. Under specific circumstances incubation time was extended to even 13 days. Every experiment was repeated at least twice. Besides homozygous and heterozygous mutants CpRI and CpL2 strains were applied as controls.

Growth curve

Synchronised cultures grown in YPD were harvested and counted as described above. 106 cells (suspended in YPD) were inoculated in 5 ml preheated (37°C) YPD supplemented with 100u/μl penicillin-streptomicin and shaken at 37°C. 100 μl were taken from the suspensions hourly (from 0 to 12 hour) and 3 hourly from (12 to 24 hour), diluted, plated onto YPD plates in triplicates and incubated for two days at 30°C. Colonies were counted and CFU/100 μl was calculated. The MAD growth curve is displayed in Figure 2.4 (C).

Strandedness

Forward and reverse cDNA primers specific for lncRNA4 transcript and a qRT-PCR primer pair were designed (see primer list). CDNA synthesis was carried out by using ThermofisherTM RevertAidTM First Strand cDNA Synthesis Kit according to the manufacturers instructions for cDNA synthesis with region specific primers. As a template RNA of *C. parapsilosis* CLIB WT isolated from phagocyte interaction was used. For both forward and reverse cDNA primers three reactions were set up individually,

two cDNA synthesis reactions containing either RNA (500 ng/reaction) or genomic DNA (0.5 ng/reaction) and one PCR with RNA template. Additionally one more PCR control with genomic DNA was applied with both the cDNA forward and reverse primers. QRT-PCR was performed with ThermofisherTM MaximaTM SYBR Green qPCR Kit in triplicates. One μl of each of the above mentioned reaction products was used as a template, and RNA (25 ng/reaction) and genomic DNA (0.5 ng/reaction) were included as controls. Melting curve analysis was performed to verify amplicon uniformity. For relative strand preference the average Ct values of RNA+cDNA_FOR and RNA+cDNA_REV cDNA synthesis reactions were compared.

Synonymous and nonsynonymous SNPs

occurrence of synonymous and nonsynonymous SNPs was observed between the Sequence of MAD1 and the blast hit in the closest related species of Candida orthopsilosis. Translation from codons to amino-acid was carried out using Biopython. The Chi square test for independence of observations was carried out in the built in R function. A p value of < 0.05 was used for significance. We validated the observations using codeML from the PAML 4 software suit (Yang, 2007).

3 ASEbyBayes a high precision software for the detection and quantification of allelic-specific expression from RNAseq data for nonmodel organisms

ASEbyBayes a high precision software for the detection and quantification of allelic-specific expression from RNAseq data for nonmodel organisms

3.1 Abstract

Motivation: In diploid organisms, two alleles of a given gene can be transcribed. Differences in expression between alleles are common, and can vary across conditions or tissues, which may have physiological consequences if the transcripts are not identical. Thus, there is a need to properly assess Allele Specific Expression (ASE). Although several pipelines do exist for quantifying ASE from RNA sequencing data, most rely on a phased reference genome. This hampers analyses of non-model organisms, strains, or cell lines without an available phased reference. Results: Here we present ASEbyBayes, an open source software for ASE analysis from RNA sequencing data in the absence of a phased reference genome. ASEbyBayes uses a two step procedure, in which single nucleotide polymorphisms (SNPs) are first called on an un-phased reference genome and then replaced by undetermined nucleotides (Ns) to remove bias in a second mapping step. Then coverage of expression, conservation over replicates, and SNPs impact on the amino acid sequence are factored in a bayesian approach for Allele Specific Expression analysis. ASEbyBayes is provided as a stand alone executable, enabling robust and reproducible analyses.

3.2 Introduction

Diploid organisms carry two distinct versions (alleles) of each autosomal locus, which can be transcribed at different levels. Differences in the level of expression between alleles may vary across tissues or conditions and have important physiological implications. This situation can be extended to polyploid organisms or duplicated regions of the genome, or even to the analysis of mixed cultures, where more than two alleles can coexist. The quantification of allele-specific expression relies on the ability to distinguish the transcript from each of the alleles. Although traditional analyses have relayed in the use of allele-specific primers or micro-arrays, nowadays most of the studies are based on high-throughput RNA sequencing (RNAseq).

Allele Specific Expression analysis often relies on single cell transcriptomics (Jiang et al., 2017; Sigurdsson et al., 2016). Established approaches use GATKs ASEReadCounter (Castel et al., 2015). Alternatively, ad hoc methods, such as binomial exact tests are in use (Bell et al., 2013). RNAseq data contains a vast amount of information, which is increasing as technological developments improve quality and quantity of individual reads. This enables more advanced analytical methods. Most existing methods for ASE rely on the availability of a phased reference genome, i.e. in which the sequences of the two allelic variants are known. This limits the application to organisms or cell lines where a phased reference genome is available. To circumvent this limitation, we developed ASEbyBayes, a bayesian approach that exploits information in RNAseq reads to derive ASE in the absence of a phased reference genome. Using a controlled simulation, we show that ASEbyBayes is superior to the commonly used GATK pipeline. Unlike available software solutions for single cell transcriptomics (Jiang et al., 2017; Sigurdsson et al., 2016), ASEbyBayes is more robust towards contaminations from subpopulations, and population derived errors in coverage.

3.3 Methods

A unique capability of ASEbyBayes is its ability to work in the absence of a phased reference genome. To reduce mapping based errors for the quantification, a two step work-flow is used. In a first step, ASEbyBayes performs SNP detection using a standard BAM file as an input. These SNPs are then replaced with unspecified nucleotides (Ns) in a new reference sequence. In a second step, reads are mapped again, this time against the modified reference. This procedure reduces possible errors derived from the bias in mapping the individual alleles. That is, in a non-phased reference reads from the alternative allele will be mapped less efficiently, resulting in a significant mapping bias that can reach up to 10-15% (Degner et al., 2009). In a subsequent phase ASEbyBayes uses an approach comparable to other downstream RNAseq data analysis such as the assessment of differential expression between genes by DESeq2 (Love et al., 2014), where distributions of read counts are modeled and then used

to find outliers (i.e. differentially expressed genes). In our approach, the expression of the individual alleles are first transformed into ratios. Then, both depth of coverage and ratio of allele expression are used to generate expected distributions. Priors are estimated from the overall mapped data over replicates, using a beta distribution, the conjugate distribution to the Negative Binomial used by DESeq and others. The expected ratio of unbiased expression for hypothetical bi-allelic SNPs in a diploid organism is 0.5. Differentially expressed alleles should significantly diverge from this ratio. Analyzing SNPs individually, compared to whole genes, gives us the ability to perform an ASE analysis on a one base resolution. Each bi-allelic SNP expression is assessed independently on whether it significantly diverges from the expected distributions. For each gene or exon, a new distribution is fitted for the SNPs that diverged significantly providing an estimate of the expression ratio for each of the alleles of the gene. ASEbyBayes exploits information derived from independent replicates by processing all replicates in the same run and using SNPs existence over replicates as an important criterion for distinguishing noisy from low-level expression. The analytical suit of ASEbyBayes is built around the htsjdk of the broad institute (http://samtools.github.io/htsjdk/). The result of an ASEbyBayes analysis is the list of observed SNPs in a Variance Caller Format (vcf) file. Additionally, a list of genes with their association of allelic expression and the respective evidence is returned in the form of a tsv file. We benchmarked our software against the most commonly used pipeline: GATK (McKenna et al., 2010), which contains SNP callers that are originally designed for DNA sequencing data, but are commonly used in RNAseq data analysis as well. We used the GATK ASEReadCounter tool to evaluate Allele Specific Expression. We used FluxSimulator (Griebel et al., 2012) to generate RNA sequencing data based on the references of the *Saccharomyces cerevisiae* chromosome 4 for a subset of 100 genes of a length between 600 and 4500 nucleotides. We generated a second reference sequence with 20 SNPs per gene in order to simulate the second allele of a diploid organism. To maintain biological impact into the simulation we estimated the ratio of synonymous versus nonsynonymous SNPs according to the published Ka/Ks ratio of 0.11, as observed in genes between *Saccharomyces cerevisiae* and *Saccharomyces mikatae* (Kellis et al., 2003). The sequencing data obtained from the

two references was combined in fixed ratios to simulate Allele Specific Expression. We used a read length of 50bp. Details on the Benchmark data, and a generator script can be found on the github repository. The source code, and the compiled software as well as the simulated data are available on the github page : https://github.com/Gabaldonlab/ASEbyBayes. Our approach showed a more robust, and significantly improved precision compared to the GATK based pipeline (See Figure 3.1). When simulating continuous allelic expression over the whole sequence (i.e. all genes show equal differential allelic expression), both the GATK pipeline and ASEbyBayes show high precision (99.8 $\pm$ 2.0 % ASEbyBayes, 99.5 $\pm$ 4.6% GATK/ASEReadCounter). However, in more realistic settings, were only a small fraction of genes showed ASE, the precision of GATK was 78 $\pm$ 5.0% while ASEbyBayes showed a much higher precision of 92.2 $\pm$ 5.1 %. Thus, our analysis suggests, that the HMM based classification of GATK tends to call False Positives in high variance environments (e.g from expression changes), while the prior based classification is more stable. False Positives are especially important in the analysis of Allele Specific Expression, since falsely called SNPs, e.g from sequencing errors, once called are indistinguishable from Allele Specific Expression. Such errors are common in population transcriptomics, and can falsely lead to wrong claims of Allele specific Expression. https://github.com/Gabaldonlab/ASEbyBayes.

3.4 Conclusion

We have developed a new software solution to quantify allele-specific expression at nucleotide resolution. We benchmarked SNP calling against the GATK Haplotypecaller and Unified Genotyper. The precision of ASEbyBayes was considerably higher, due to the overestimation of SNPs in areas of low coverage by the GATK approaches. These false positives would potentially skew the downstream analysis towards false conclusions.

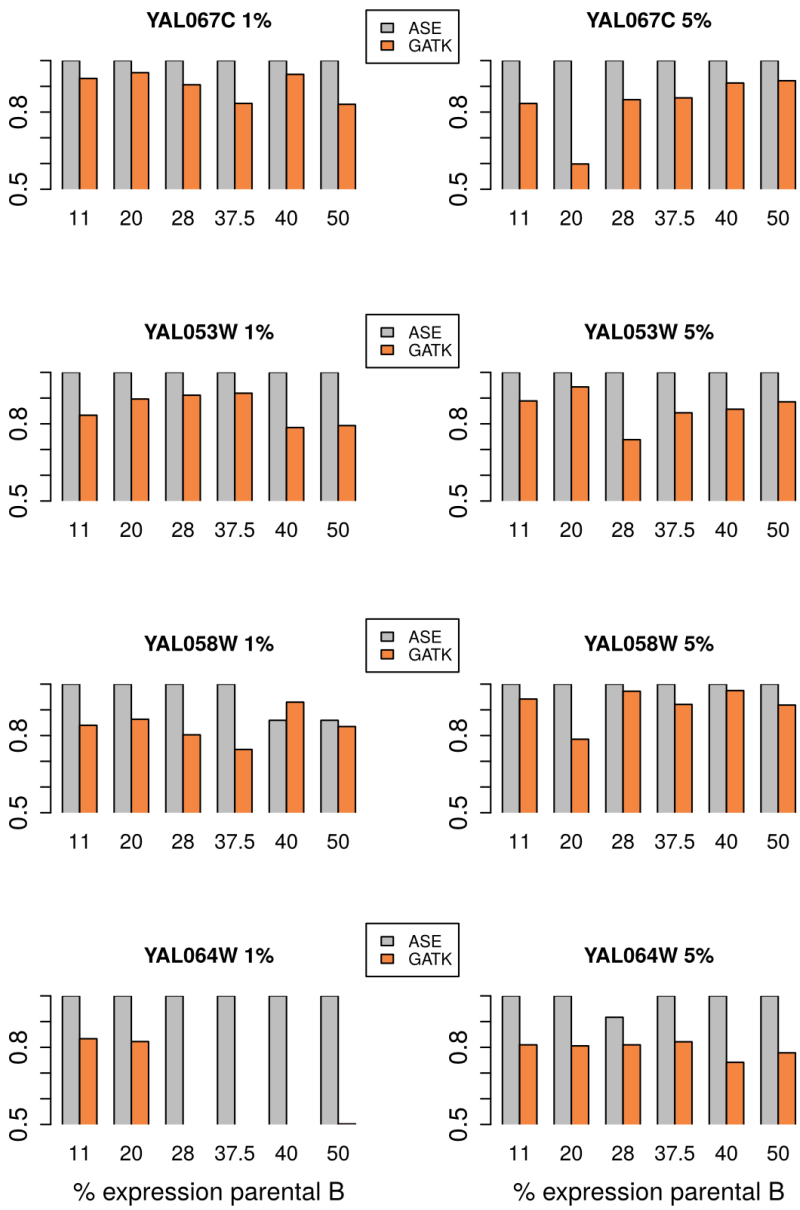

Figure 3.1: Benchmark. Figure showing the Benchmark of ASEbyBayes against the GATK pipeline. Simulated RNAseq data was produced showing only Allele Specific Expression in single genes. 4 independent genes are shown. The left side contains a 1% divergence against the reference, the right side shows a 5% divergence. Notably, there is no linear correlation of the GATK error against extend of divergence.

3.5 Methodology and Benchmark

Methodology

The methodology followed in ASEbyBayes was modeled after the GATK (McKenna et al., 2010) approach for SNP calling, but improved classification for the more complex noise background of nonmodel population transcriptomics. The software takes several BAM files of biological replicates of already mapped input RNAsequencig data. SNP calling is performed internally by the software, in order to preserve the information necessary to estimate noise in the subsequent analysis. The SNP calling is based on the htsjdk Locuswalker, and modeled according to the samtools mpileup (Li et al., 2009) approach of evaluating coverage for non-reference nucleotides [see source code // BAMHandler]. SNP coverage is considered as a ratio between the respective alleles. The respective likelyhood distribution is modeled according to a Beta distribution i) (Raiffa and Schlaifer, 1961), the conjugate distribution to the negative binomial, used e.g by DESeq2 (Love et al., 2014) to model RNA sequencing reads. The beta function with corresponding likelihood function is used to generate the posterior density. We relied on conjugate distributions in the analysis to avoid the varying impact of the analytical integration problem posed by the application of Bayes law (Fink, 1997)

$B(p,q) = \frac{((p-1)!*(q-1)!)}{(p+q-1)!}$
i) Beta distribution

$pdf(x) = \frac{(x^{(p-1)}*(1-x)^{(q-1)})}{B(p,q)}$
ii) Probability Density Function for Beta distribution

To estimate the variance of Noise expression, replicates are compared. Low coverage non-synonymous SNPs not reproduced in replicates are considered the baseline for noise following the above described distribution, and are used to estimate the prior for the subsequent noise hypothesis. If no replicates are given, a threshold is introduced to limit the impact of

false positives however. It is not recommended to use the application without replicates. Probability density function of SNPs as described in ii) are considered for subsequent classification. Notably, for the Noise classification, an additional weight for synonymity corresponding to the inverse likelyhood of natural occurace is considered. This corresponds to the assumptions that real SNPs are more likely to be synonymous due to a lack of purifying selection. In the next step, a new reference FASTA containing all non-noise SNPs masked (replaced with 'N') is returned. At this point, a remapping of the raw data is recommended to improve quantification, and remove a bias resulting from similarity of one allele to the un-phased reference. After remapping, the SNPs called in the first step are quantified. Priors assuming an equal allelic expression are generated as base priors. The pool of hypothesis is extended by central limit until no further hypothesis are needed to explain the data. Alpha beta values are adjusted according to the respective hypothesis and the total expression of the gene. Since the posterior distribution derived from a conjugate prior belongs to the same distribution, the probability density function of each SNP is equally generated according to ii). Ultimately, the analysis targets the expression of whole genes. An important consideration for the conclusions on whole gene expression by individual SNPs is the heterogeneity of SNPs origins on genes. Several SNPs on a single gene may derive from different sources. Potential origins considered are allelic expression or subpopulations as well as sequencing errors. Contamination of the RNA sequencing data by subpopulations is analyzed by comparing outliers of individual SNPs against the majority hypothesis on genes over the whole sample. The probability of each SNP to correspond to either active or passive Allelic expression is estimated by evaluating its probability for each function considering the cumulative probabilities for all hypothesis according to Bayes theorem. Genes are classified according to their majority function, the hypothesis containing the highest cumulative probability density. Potential contamination SNPs are noted. Subsequent steps of development will make use of the contamination marked data, to infer the likelihood of contamination. Additionally, frameworks for the analysis of polyploidy are being developed.

Benchmark

GATK is the leading platform for genome analysis. Its creators described its usage in variance calling from RNAseq in the following tutorial : http://gatkforums.broadinstitute.org/gatk/discussion/3891/calling-variants-in-rnaseq. We decided to use the GATK toolkits HaplotypeCaller and ASEReadCounter as benchmark software. Tophat2 (Kim et al., 2013) was used for mapping, since it is capable of introducing read groups required by GATK. We applied Flux Simulator (Griebel et al., 2012) to simulate RNA sequencing reads. The simulations were based on *Saccharomyces cerevisiae* chromosome 1, using only the first 100 genes. References for heterogeneity were generated at 0, 0.25, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 4.0 and 5.0 % (mutations against the reference). We generated various degrees of coverage, 20, 60 100, 150, 250 and 500.000 reads for the samples. Homogeneous Allelic expression was introduced by combining ratios of various degrees of heterogeneity. Both softwares behaved nearly identically precise. We therefore did not plot the results. Due to the lack of replicates used, in following the GATK pipeline, ASEbyBayes estimates fewer SNPs than GATK, but maintains a low rate of false positives. FluxSimulator does not allow the introduction of varying expression. We therefore introduced variation by only introducing SNPs into specific genes. References were generated, in which on of five simulated genes YAL048C, YAL053W, YAL058W, YAL064W, YAL067C were mutated to 5% of their sequence. Similar to above, coverages were generated. In these datasets, GATK performed substantially less stable than ASEbyBayes. As mentioned above, the most likely explanation is the lack of training data for GATKs Hidden Markov Model (HMM). ASEbyBayes is more cautious on SNP calling due to its priors, GATK seems to overestimate SNP density. GATK recommends a downstream application, MAMBA, to evaluate Genewise expression. After contacting the authors without a response, we found the application to be no longer in active development, and requiring unspecified formats for processing, the downstream analysis software was therefore omitted. The observed ratio of False positives predicted by GATK establishes varying amounts of uncertainty. The amount of misclassification for potential Allele Specific Expression depends on the Thresholds used. The given precision of $78 \pm 5\%$ suggests that up to 22 of

100 SNPs are misclassified, potentially expanding the predicted amount of Allelic Expression. In our benchmark, due to the random distribution of False positives by GATK, 2 to 6 genes were estimated to contain Allele Specific Expression, while only one gene contained true SNPs, resulting in a rate of false discovery of up to 6:1 by GATK. ASEbyBayes found low amounts (1-3 SNPs) outside the mutated genes, and due to the Bayesian inertia, did not predict false positive genes. Notably, the produced data oversimplifies the complexity of real samples. Additionally, following the GATK approach we used both softwares without replicates, which is not recommended for ASEbyBayes. Additionally, no remapping was performed to generate output similar to GATK. Due to the design of ASEbyBayes, it is expected that its SNP calling and quantification quality improves if replicates are provided and re-mapping is performed.

4 Investigating cancer derived Monocytes THP-1 capabilities in pathogen research via comparative transcriptomics

Investigating cancer derived Monocytes THP-1 capabilities in biomedical research via comparative transcriptomics

4.1 Abstract

Undifferentiated human monocytes encounter various pathogens while present in the bloodstream. They are considered a primary responder and regulator for human immune reactions. As such, experiments investigating responses to pathogens are often reliant on Monocyte cell cultures. For reproducibility reasons, immortalized cell lines are used. One of the most important cell lines used to model pathogen interactions is THP-1, which has been used in a variety of high throughput transcriptomics experiments. Yet as a cancer derived cell line it may no longer maintain its orignial functionality in detecting and responding to pathogens. Using available large scale transcriptomics datasets, we investigate the comparative response of THP-1 to a variety of human pathogens; viruses, bacteria, protozoa and fungi. Our approach focuses on the behavior of THP1 in its response to the different pathogens. Our aim is to provide comparative insights into the cell lines behavior and capabilities to potentially improve future experimental design.

4.2 Introduction

Undifferentiated human monocytes reside in the bloodstream for up to three days before differentiating and moving into tissue. During this time, they are a primary responder to any invasive pathogen entering the bloodstream. Due to their primary role in coordinating host responses, they have been suggested as targets for immune augmentations strategies e.g during fungal infections (Segal, 2007). Most experiments investigating bloodstream infections to pathogens rely on the use of immortalized cell lines to reliably and reproducibly model potential interactions between humans and pathogens e.g (Leland and Ginocchio, 2007). A potential downside of such cell lines, usually derived from cancer lines, is that they

are established after human cells have already mutated to a very unnatural cell state (Kaur and Dufour, 2012).

A common cell line used to study the behavior of Monocytes is the leukemia derived cell line THP-1 (Tsuchiya et al., 1980). This cell line has been used to study a variety of pathogens using RNA sequencing based transcriptomics. Experiments using THP-1 interaction models involve interactions with viruses, Ebola and Marburgvirus (Martinez et al., 2013), Zika (Hanners et al., 2016) and bacteria such as *Coxiella burnetii* (Millar et al., 2015), *mycobacteria spp*. (Reyes et al., 1999; Zakharova et al., 2010) as well as the protist *leishmania mexicana* (Millar et al., 2015). In a recent study, Toth et al. (Toth2017) Investigated the response of the pathogenic yeast *Candida parapsilosis* to THP-1. Additionally, data for interactions to the chemicals ethanol and calcitriol is available, as well as a compound called Tissue-type Plasminogen Activator (TPA), which causes differentiation to macrophages (Barendsen et al., 2008). To our knowledge, no direct investigation into the comparative response behavior of the cell line THP-1 has been carried out so far. In this study we hope to provide insights into the behavior of THP-1 if exposed to different human pathogens, and evaluate its ability to develop specific responses. To address how THP1 cells respond to the different stimuli, we investigated transcriptional profiles via RNA sequencing for the human THP1 cell line after exposure to the above mentioned pathogens and chemicals. To reduce analytical bias from the individual experiments, our analysis began with raw RNA sequencing data available at the NCBI Sequence Read Archive for the individual projects. A common, standardized pipeline was applied to carry out the data processing. We focused the subsequent analysis on the inter-project response for the individual pathogens. In order to overcome the very different experimental setups, we relied on more global approaches for tanscriptomic analysis. An important focus was the ability of the cancer derived THP-1 cell lines general response to distinguish the individual pathogens. Specific responses have been investigated in the individual experiments, yet such a variety of pathogens is expected to trigger substantially different overall response pathways. We investigated the impact of the individual pathogens via dimensionality reduction based clustering, and comparative GO enrichment. Both on large

scale to compare the overall behavior of the cells, and on the two available time course analysis to investigate the more minute temporal dynamic of transcription shifting. The available time course analysis comprise the bacterium and intracellular pathogen *Mycobacterium abscessus* and the yeast *Candida parapsilosis*. The *mycobacteria* species consist of a range of bacteria best studied for causing tuberculosis. They are documented to be fast growing and potentially multidrug resistant. The *M. abscessus* complex is also resistant to disinfectants and, therefore, can cause postsurgical and postprocedural infections (Lee et al., 2015). Candida species cause common nosochomial infections (Casadevall and Pirofski, 1999). As yeast, their mechanisms of pathogenicity differs significantly from that of bacteria. The yeast potentially inducing a much weaker response, testing the limits of THP-1 transcriptome adaptation.

4.3 Material and Methods

RNA sequencing data processing

RNA sequencing data was obtained in its raw .sra format from the Sequence read archive (Leinonen et al., 2011), with the exception of the data for *C. parapsilosis*, which was provided by the authors of Toth et al. (Tóth et al., 2017), the full list of sequence runs can be found in the supplementary table 1. A shell script to initiate the full download can be found at the github repository https://github.com/Gabaldonlab/THP1data . After extraction using the sratoolkit. Trimming, for quality pre-processing, of the reads was performed via Trimmomatic v0.32 (Bolger et al., 2014). We mapped the reads using the STAR (Dobin et al., 2013) mapper against the hg38 human reference genome. Human genomic data hg38 v 81., both the reference and annotation files were obtained from UCSC (Ucsc and Browser, 2003). Reads were counted using the htseq package (Anders et al., 2015). For the analysis only annotated exons were considered. An overview of samples considered is presented in Table 4.1.

Pathogen	Sample count	Time course	Quality control	Archive
Mycobacterium bovis	14	No	failed	ERR5604
Mycobacterium abscessus	19	Yes	pass	SRR23160
Zika	2	No	pass	SRR51901
Ebola	2	No	failed	SRR16602
Marburgvirus	2	No	failed	SRR16367
Leishmania mexicana	4	No	pass	SRR1562
Coxiella burnetii	4	No	pass	SRR1562
Candida parapsilosis	16	Yes	pass	from author
Staphylococcus aureus	9	No	failed	ERR50285
Non Pathogen	5	No	pass	SRR16365

Table 4.1: Table of RNA seq runs processed in the analysis. Lists numbers of samples, pass of the mentioned quality check and accession number their archive location.

Data processing

Read count normalization was performed via transcript per million TPM normalization. R libraries were used to investigate the Principal Components underlying the data variability. The R built in prcomp module and the library FactoMineR (Le et al., 2008) were used to compute the PCA and cluster estimation respectively. Tree based hierarchical clustering was carried out using the python scipy library. Gene enrichment was analyzed via python scripts available on the projects github https://github.com/Gabaldonlab/THP1data , generating a background model of variance. The enrichment compared to the full human background was carried out using the GOrilla tool (Eden et al., 2009)

Visualization

Visualization was performed via the R module ggbiplot, based on the ggplot2 library, as well as the FactoMineR and superheat plotting function for the respective R scripts. Visualization in python was produce via matplotlib.

Enrichment analysis

Expression enrichment for unregulated genes was computed for each gene on the variance over the non pathogen derived conditions, uninfected cells and separately against the chemicals ethanol and calcitriol. Outliers were tested against a normal distribution using student t-test. The method described by Benjamini & Hochberg (Benjamini1995) was used to correct for multiple testing and evaluate false discovery rate (FDR), in order to correct the resulting p-values. Adjusted p-values of < 0.05 were considered significant and analyzed by GOrilla against a total background.

4.4 Results and Discussion

As shown in Figure 1 the Principal Component Analysis showed a clear and distinct response to the individual conditions. This can be considered an important sign that the THP1 cell line has retained its potential for detection of the individual pathogens in initiating individual responses. In a PCA, the abstract underlying effects are quantified and shown on components or axes, with relative strength per axis denoted in percent of variance explained. Individual principal components can show multidimensional response factors. Dimensions are visualized in Figure 4.1. Due to the complexity of analysis, the first four dimensions were considered to explain sufficient variance, collectively accounting for 36.5% of the observed variance. Similar responses were observed between the intracellular bacterial pathogens *Coxiella burnetii* and *Mycobacterium abscessus*, derived from experiments performed by indpendent groups [supplementary table 2], suggesting that the first components are not influenced by the sequencing but directly by the THP-1 response, the profile of the two strains diverges in the third component showing more nuanced differences in response of THP-1 between the two pathogens.

Overall four distinct response clusters can be observed. With separation of clusters occurring for virus to yeast in the first component, and a distinction of bacteria over the second and third. The intracellular *M. abscessus* and the yeast *C. parapsilosis* are the most robust groups due to a larger sample size of 19 and 16 runs respectively, and replicates over

a time course of infection assay. Factorial analysis to investigate time point responses was therefore limited to those two species. To gain a more detailed view on the minute behavior we investigated the response to the two larger time course analysis. Data was produced for rough and smooth morphologies during exposure times of 1, 4 and 24 hours. Although the analysis shows a clear separation (see Figure 4.2), the primary component derives from the effect between the replicates. The response to *C. parapsilosis* was less homogeneous, most likely due to the lack of replicates and the overall lower pathogenicity of *Candida* as compared to *Mycobacteria*. In the next step, we quantified overall transcriptional responses against background noise models visualized in Figure 4.3. Two Noise models were designed. In the first, (background) we used the average counts per gene in uninfected samples, to evaluate the expression compared to an uninfected baseline. For the second model (stressors) we evaluated the average gene count for samples treated with TPA, ethanol and the Vitamin D metabolite calcitriol to evaluate basic stress responses. Individual genes were tested for upregulation only against the model genes, returning a value of significance per gene and pathogen. This methods ignores experimental design in order to generalize the various experiments. Figure 3 shows the overlap between the background and stressor comparison. In total, 9302 genes were activated in any pathogen response compared to both backgrounds, with 1318 genes overlapping between the noise models. 6780 and 1204 were unique to the background and stressors model, respectively. This indicates an active response by THP-1, and its ability to distinguish uninfected surroundings to chemical stimuli.The cells show a clear distinction between the responses to the individual pathogens. Yet, the THP-1 response to *C. parapsilosis* shows no significant difference to the background model, but does present a distinction to the stressors model. *C. parapsilosis*, as a pathogenic yeast, is often commensal e.g (Gabaldon et al., 2016), and seemingly does not provoke a general strong transcriptomic response in a naturally commensal host. Interestingly, compared to the stressor background, only the *M. abscessus* cells show a classified defense response. According to GO terms, the response is antiviral [Supplementary table2]. An antiviral response to *M. abscessus* is partially expected, due to the related pathogens *M. tuberculosis* ability to trigger Interferon (Prabhakar

et al., 2003), and Interferon production in general T-cell reponses (Belardelli and Gresser, 2010). *L. mexicana* shows the strongest overall response. With 3861 more genes activated than in the background, and a response of 694 unique genes to the *L. mexicana* stress compared to the stressor background. GO enrichment using GOrilla visualized in table 1 shows unique responses for the pathogens. Most notably, the most enriched GO term for Zika is GO:0030900, pertaining to anatomical structure and forebrain development. This is in line with Zikas clinical symptoms, such as microcephalus as described in this review (Paixo et al., 2016). Enrichment for *leishmania* showed active genes involved in iron transport, an observation in line with Huynh et aliis (Huynh et al., 2006) discovery of iron transporters being essential for parasitic reproduction. Active Iron transportation could therefore be an expected host response.

4.5 Conclusion

Our analysis suggests that the THP-1 cell line is capable of distinguishing various cellular stresses, and provide individual responses to various chemicals and pathogens. It accurately portraits gene enrichment for e.g Zika clinical symptoms. The large scale transcriptomic response is uniquely different for each analyzed experiment, and the cells show similar responses to the intracellular pathogens e.g *Coxiella burnetii* and *Mycobacterium abscessus*. Yet, the experimental resolution is less pronounced in more detailed experiments. Time course analysis using THP-1 showed a stronger variation between technical replicates than the actual experimental course. A recent study by Schurch et al (Schurch et al., 2016) estimates the number of true positives in RNA sequencing with 3 replicates to be between 20 and 40%. Using the first components as indicators of variance, we estimate that at least 12% of the total variance are attributed to technical variation. By using general probabiliy, we can estimate the true positives for triplicates to be between 13.6 and 27.26% using THP-1 cells (according to $(1-(1-\alpha)^n)$). Therefore, especially for the investigation of pathogens with an expected mild response, additional replicates are strongly recommended.

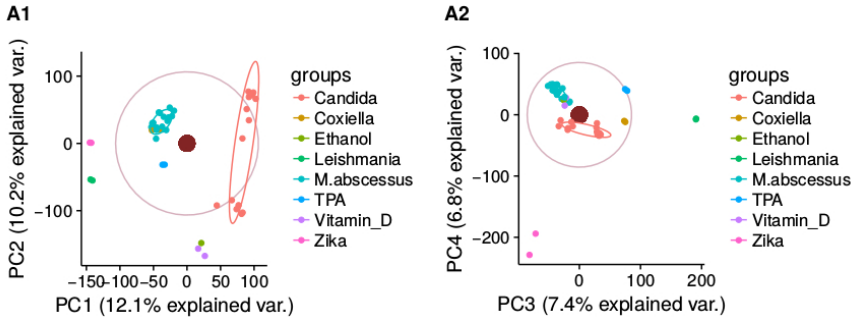

Figure 4.1: Principal component analysis for all genes, normalized to TPM. Due to the relative low variation per component the first four dimensions are displayed in two 2dimensional plots. Components are displayed in two plots PC1 and 2 in plot A1 and PC3 and 4 in A2. In A1, a cluster separation between the different pathogens and stressors can be observed on the first component. Two clusters comprised of the timeline experiments on *M. abscessus* and *C.parapsilosis* are visible along the first axis. A2 seperates viruses more clearly from the protist, as well as the intracellular bacteria *Coxiella bruneii* and *M. abscessus*.

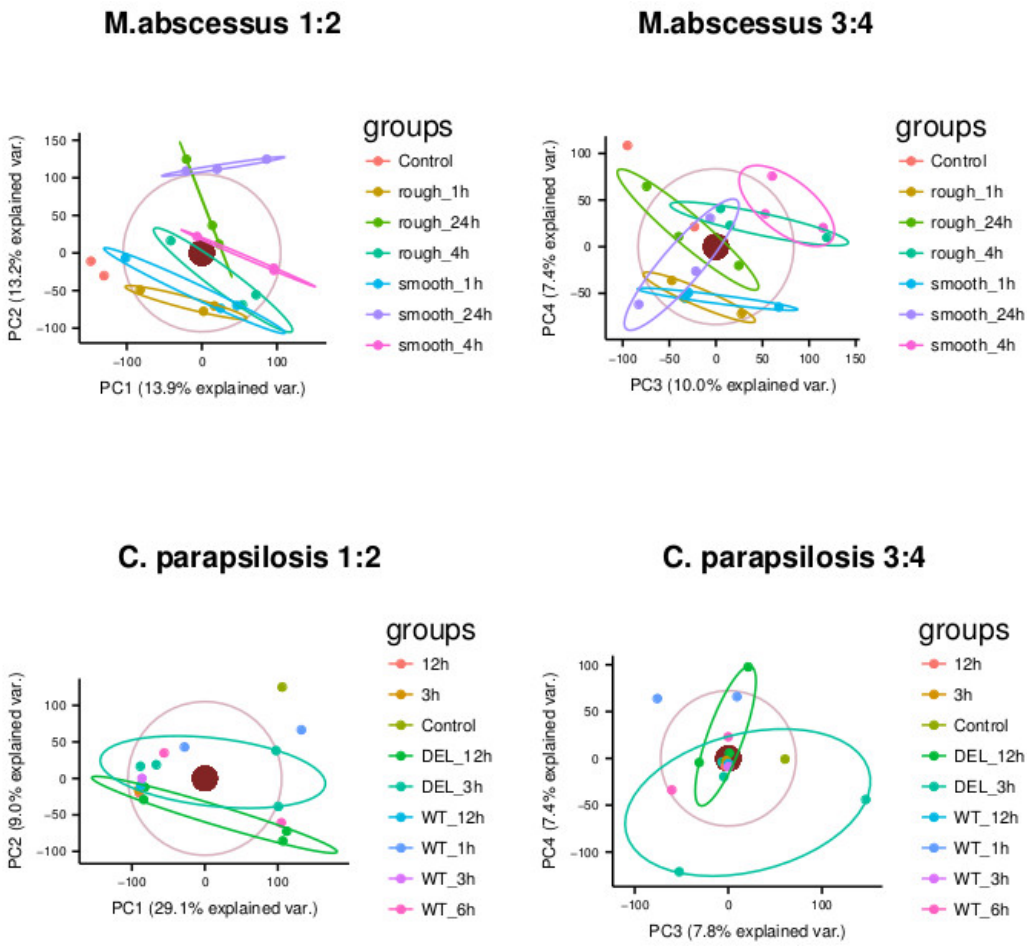

Figure 4.2: PCA of the time course analysis for the infividual sets of *M. abscessus* (above) and *C. parapsilosis* (below). Both studies show variance between technical replicates to be responsible for their first component, suggesting a lack of regulation by the THP-1 cells, or rather a lack of designated response. While *M. abscessus* shows a distinguishable variability over the time course, clustering on *C. parapsilosis* does not separate the individual conditions.

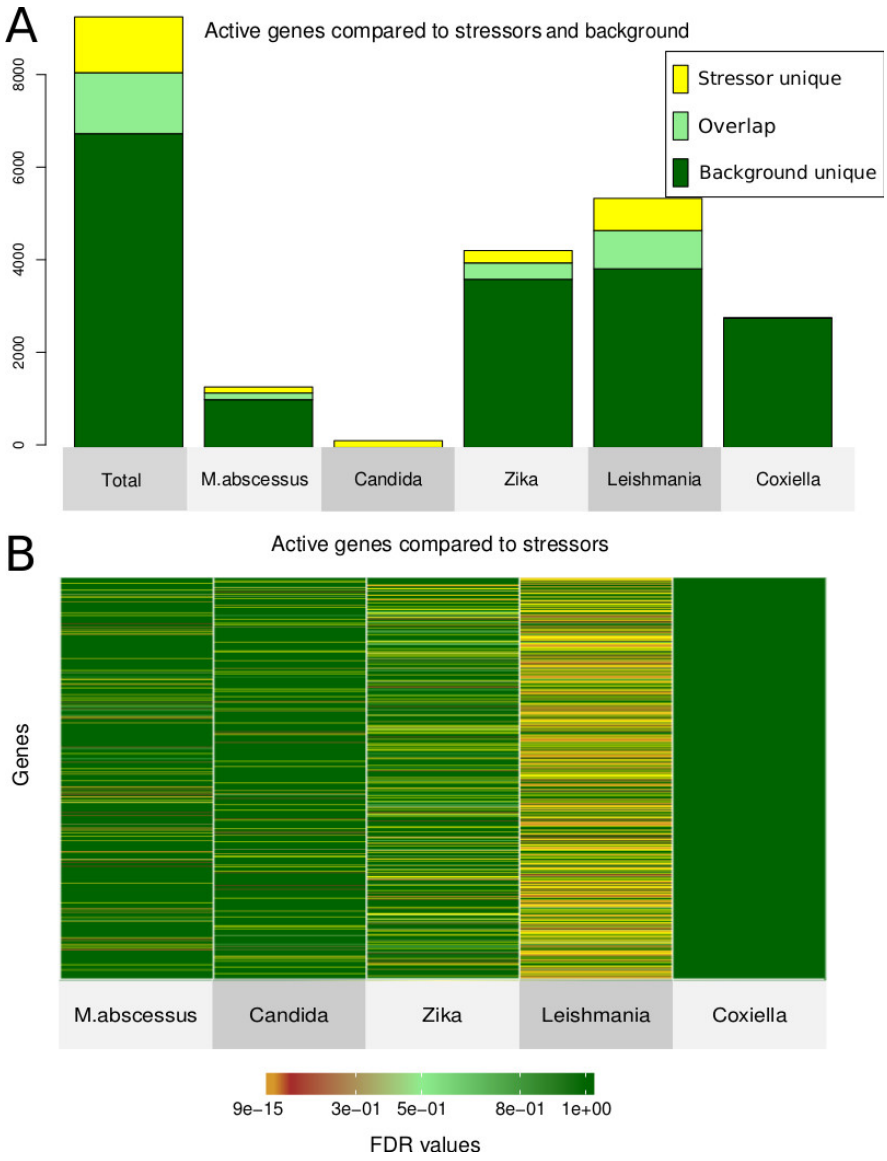

Figure 4.3: (A) Barplot visualizing the quantification of response against the two background models. Unique responses in yellow and dark green are only observed against the specific background. Overlap in light green shows responses similar between stress and uninfected cells. (B) shows a heatmap of individual genes actively regulated in the 5 species against the stress response. Notably, *Coxiella burnetii* shows no significantly enriched genes against the stressor subset.

5 General discussion

As mentioned in the introduction, bioinformatics is a young field. The applications and approaches are still actively developing. We are only beginning to explore the true potential of the new technologies and their capabilities. All the projects described above rely in some way on the analysis of RNA sequencing data. Different projects capture different aspects of potential analysis pathways. RNA sequencing is most commonly used to evaluate intracellular behaviour in cells under stress conditions. So most analysis carried out via RNA seq to some extent resemble the projects described above concerning the interaction analysis of THP1 - *Candida parapsilosis* or the analysis of Allelic expression. Due to the engagement in the form of a Marie Curie Intensive Training Network, many collaborations in the analysis and evaluation of RNA sequencing data were established. Using the pipeline described above standardized the projects of the consortium. The core projects described concern deeper investigation into the capabilities of RNA sequencing. Three independent approaches are elaborated in the three respective chapters. The most extensive project deals with the downstream analysis of transcriptomics. Investigating long noncoding RNAs from RNA sequencing is an established approach in higher mammals, but not in nonmodel species. Additionally, we inferred functionality using advanced clustering methods. In the other projects we looked both at the analysis for complex interaction models via RNA sequencing, and the establishment of more complex analytical pipelines.

Modern transcriptomics is a powerful methodology, resulting in vast amounts of information, but analysis of anything but the most basic setups can be difficult. Especially interaction transcriptomics, e.g between pathogen and host, still pose real difficulties in data analysis. The problems derive

from noisy populations and heterogeneous responses as well as secondary effects on expression caused by unmeasured agents. Such secondary effects are especially pronounced in more complex model organisms.

The projects described above also commonly investigate the behaviour of pathogenic yeast. Due to the alternative concept of *Candida* pathogenesis compared to other pathogens, the analysis of *Candida* human infection models is a complex undertaking. Many individual responses of the yeasts form a combined invasive response. Although our setup included time courses to capture shifts in responses, the obtained noise levels were initially discouraging. The problems of time course analysis derives from the nature of infections. Infections are not a binary process that either does or does not occur. In fact, they occur transiently over a time frame, with gradually shifting responses by each participant. Our approach to investigate undifferentiated Macrophages shows an active response that diverges from the ones observed in *C. albicans*.

The project used relatively early RNA sequencing data, with protocols designed before 2013. Older protocols were more noisy in general and less consistent in coverage than current RNA sequencing. Additionally, the analysis was performed on populations of data. Alternative Methods, such as SAGE (Vilain et al., 2003) can already be used to analyze single cell transcriptomics. While RNA sequencing gives a good estimate of overall expression, investigating the responses of single cells is a very recent, and not yet widely used approach. For modern RNA sequencing a minimum quantity of RNA is required, Illumina sequencing requires 1-10μg. The amount is lower for PCR based approaches like Illumina. This can be achieved in human cells, but not yet in *Candida* species.

In order to overcome this challenge, we carefully adapted the pipeline to accommodate the noise. We increased the expected activation threshold for the transcripts, and added an additional differential expression estimator. We generalized the analysis to rely more on the gene set enrichment, a measurement more general than Differential Gene Expression (Conesa et al., 2016). The stability of enrichment analysis comes from the concept, that

random noise measured will act equally strong on any gene, and will therefore introduce noise without activating whole pathways. Analyzing the data via Gene Ontology enrichment focuses the analysis entirely on pathways, and therefore direction. Finally, we managed to generalize the response to a degree allowing us to establish a comparable analysis between *C. parapsilosis* and a more general band of other common human pathogens on the same human cell line. The inverse project, the description of the human response, formed the background of the analysis presented in Chapter 4.

Since Next Generation Sequencing approaches were only developed in the last decade, their range of applications is not yet fully explored. Apart from the more common Differential Expression analysis we investigated the analysis of Allele Specific Expression. To overcome the lack of reliable software solutions for our specific interest, we had to establish our own approach. The results of this project so far are mostly methodological. Available software suits target single cell transcriptomics, with a very different error model compared to yeast population. Implementing Allele Specific Expression analysis on non model organisms required an adaption to common approaches. Previous studies relied either on a combination of approaches from different software suits, or the implementation of ad-hoc testing. The result of these projects are combined in the software ASEbyBayes. The open source project allows a standardized way of analyzing RNA sequencing data for Allele Specific Expression. Clear priority was given to the analysis of nonmodel organisms, such as yeasts. Allele Specific Expression is a field in which sufficient quantified data is available to implement a classification approach that can draw upon more general knowledge. An important point is our attempt to remove mapping errors by encouraging a remapping step. With increasing quality of the newest versions of Illumina sequencing this method will become even more robust, since more and more noise is removed from the data before analysis by the sequencing itself. An important note on the Software is that its sequencing error estimation is based on the Illumina short read sequencing error distribution. For upcoming short and long read sequencers, additional evaluation is necessary. There are several ways for extending the project.

The first and foremost is the biological application of the pipeline. The initial question of the project concerned the investigation of expression in a *C. orthopsilosis* strain that is a known hybrid, as described by (Pryszcz et al., 2015), yet for which only one parental strain is known. The two parentals diverge about 5 percent at the nucleotide level. We initiated a primary analysis on unpublished data. *C. orthopsilosis* is closely related to *C. parapsilosis*, but shows a parasexual cycle. *C. parapsilosis* shows few heterozygous reads due to its lack of mating. Recent studies suggest that this hybridization in yeasts is a common occurrence, potentially increasing the target applications for the tool. One of the mayor implementations will be the introduction of the statistical framework for alloploidy. Alloploidy is defined as variable chromosome number. Given sufficient coverage, SNPs carried on one or two out of multiple alleles can be analyzed directly. An estimation of ploidy, and possibly variable effects can be analyzed. This analysis will require a shift in experimental design and statistical analysis, but is within the capabilities of the tool, given sufficient coverage. The main limiting factor for accuracy in the analysis is the lack of coverage over some regions. As the law of large numbers dictates that given sufficiently high coverage allele wise expression of SNPs should diverge towards its true ratios.

Another important step in the analysis of modern RNA sequencing data is its usage in investigating new cellular mechanisms. The recent interest in noncoding RNA research and increased availability for RNA sequencing data enable new paths of investigation. In the project described in Chapter 2, we investigated noncoding RNA in the species *C. parapsilosis*. Based on little available knowledge of noncoding RNA in other yeast species, it was necessary to adapt pipelines build from mammalian transcriptomics. By utilizing RNA sequencing data, and combining it with more advanced predictive modeling approaches we generated a subset of predicted functional long noncoding RNAs. Several features of the analysis provided interesting findings. The original number of long noncoding RNAs in the range of 300 – 600 seemed quite low, compared with higher eukaryotes. Yet the relatively low level detected may stem from the very dense genome found in yeast. Where higher eukaryotes may contain sufficient intergenic regions to en-

code their functional noncoding RNAs, yeasts may need to double down on existing genes. Our functional predictions were an application of methods developed for the study of microarrays. Using weighted gene co-expression analysis of variance stabilized data proved quite successful. Our association of noncoding features to functionality is accurate, given the expected variation due to the limited sample size and lack of annotation data. Due to the limited amount of available conditions, the resulting predictions for *Candida parapsilosis* produced an overly dense network. But nonetheless served as predictors to the later phenotype studies. Especially after the filtering through the interaction prediction of a fundamentally different methodology via the interaction predictions of CatRapid. By selecting the transcripts that not only co-express but are also predicted to directly bind to protein coding genes we hoped to reduce the number of false positives. Although the additional filter limits the potential functions of our noncoding RNAs. In our case, our predictions were removed for potential long noncoding RNAs in other potential functions. Evaluating those functionalities will be one of the further downstream analysis of this project. Notable is the remarkably high rate of observable phenotypes in the deletion mutants. With 4 of 5 knock outs actively showing a mild change in phenotypes and a weakening in pathogenicity in *galleria* models. And 3 out of 5 a measurable responses in the phenotype screening. This exceeds the amount of functional phenotypes observed by previous approaches in related yeasts e.g by Holland et al. (Holland et al., 2014). In order to focus our investigation we had to limit proceedings to one transcript. Transcript 4, called MAD1, showed the most interesting phenotype. Our analysis strongly suggests that the transcript is antisense transcribed in its function as a temperature response regulator and in the handling of Cadmium based ionic stress. This regulation may not be the only function of the transcribed locus. The reintegration of the open reading frame did show a different phenotype than the frame shifted transcript reintegration. Analysis of the secondary structure formation potential shows that the forward and reverse strand fold into a significantly similar structure. A common assumption of long noncoding RNAs is that their function derives from their secondary structure. This could be a potential explanation, since both strands fold in the same way, they may be capable of performing the same function. With the protein coding addition being an additional step

in regulation, but not a necessity. Ultimately, the phenotype studies showed that the transcripts reintegration of the only the ORF in forward, reverse and frameshift all restored the phenotype. The expression profile showed an extended transcription beyond the boundaries of the ORF. The re-integrated ORF was additionally put under a different, and much stronger promoter than the transcript would have been in a natural environment, especially given the 99.6 to 0.4 % transcription of the antisense compared to the ORF. Overall, a function for the protein coding region cannot be ruled out. The potential translated protein could convey a significant part of the function resulting in the phenotype. Disproving the function of a potential protein is challenging. Since approaches such as Mass spectrometry can only be used to prove proteins, not disprove them. And few approaches reliably enable the breakage of the secondary structure, especially since it shows on both the sense and antisense transcript.

Ultimately, the interesting phenotype of the transcript, whether protein coding or not, points to a possible regulatory function. This was unexpected from a transcript of a dysfunctional mating type locus. It was early on implicated to play a role in the pathogenicity of *C. parapsilosis*. Resistance to physiological temperature is a mayor adaptation for any yeast as a mammalian pathogen. The studies on model organisms, and transcriptomics analysis revealed certain insights into the functionality. We can estimate which pathways are actually affected by MAD1 / MADCaT. The precise mechanisms are still unknown, and will have to be researched in subsequent projects.

C. albicans can regulate 400 of its genes to introduce a morphological shift as a response to external stimuli. Discussed in the introduction as white-opaque switching. It has been argued that this shift is epigenetic, due to a lack of genomic changes. The states are propagated through positive feedback loops involving WOT1. *C. albicans* performs a switch from opaque to white colonies at 37°C. *C. parapsilosis* is not known to switch its morphology. One of the most remarkable features of the temperature sensitivity in the MAD1 deletion mutant is its shift towards sensitivity over time. The cultures behave like wild type for up to 10 hours in YPD, after which their growth slows. The changes observed in the mouse model hint to an altered cell-wall

composition that change cellular affinities. This alteration might be a result of failing epigenetic regulatory mechanism that would otherwise enable adaptation. The slowed growth could be a hint to a mechanism similar to white-opaque switching in *C. parapsilosis*. This mechanism may have so far been undetected because the morphological feature changes are insufficient for detection by classical observations. The current state of investigation into *C. Parapsilosis* does not yet allow for epigenetic analysis. A potential next step in this investigation would be Chromatin ImmunoPrecipitation sequencing (ChIP seq), to investigate chromatin changes derived from absence of the transcript.

The project initially began with the simple detection of long noncoding RNAs. But the observed transcripts, the confirmation of the phenotypes, and the recent advances in transcriptomics have opened up several new and interesting research pathways. The most promising downstream analysis involve the two other transcripts that show measurable phenotypes. Downstream validation for the individual transcripts will be extensive and consist of several independent projects.

LncRNA 3 shows the strongest phenotype of all the tested long noncoding RNAs. Overall the observed changes in function surpass long noncoding RNA 4. The deletion mutants show a number of sensitivities to a variety of stresses. Cell stress sensitivity is increased to an unstable in its absence. This provides some difficulty in reproducibility. Only one diploid deletion mutant was successfully created so far. For all other transcripts, two deletion mutants were created simultaneously. The strong phenotype hints at an important role for this noncoding RNA, but compared to the other transcripts, its noncoding RNA potential is more questionable, since it overlaps with a more robust protein coding gene. The partial overlap can be fixed, and the true effect of the noncoding RNA estimated by reintroducing that protein in full, and subtracting the effect from the observed phenotype.

The relatively weak phenotype made transcript 5 less attractive than for initial downstream analysis. But one major advantage is the lack of overlap with an longer ORF, or structural similarity to known domains. This long

noncoding RNA would be easier to prove as a proper noncoding transcript. Although the possibility of it being a carrier for translated peptides can not be ruled out either. Further study has to be undertaken in order to analyze this transcripts as a potential initial functional noncoding RNA, with a less interesting phenotype.

The CRISPR/Cas system has only recently been established for the *Candida* system (Vyas et al., 2015). For our analysis an older, and more work intensive approach was applied. With the introduction of Crispr/Cas, new deletion mutants can be created more easily, and the subsequent analysis can be done more rapidly. Five deletion mutants were created from the 17 predictions of functional long noncoding RNAs. The 17 functional ones were a subset of 64 conserved noncoding RNAs. With the demonstrated rate of observed phenotypes, the remaining transcripts provide interesting targets on their own. With the ground work established by the first five candidates, more functionalities can be investigated. Our production of the additional strand specific data will improve classifications even more. Initial analysis of the strand specific data showed large amounts of antisense transcription. Yet any claim in this direction would have to be validated extensively. Our predictions in *Saccharomyces cerevisiae* have also been very encouraging. Little is known about the involvement of long noncoding RNAs in baker's yeast. At the time of this writing only 12 long noncoding RNAs were annotated. Our initial predictions were backed up by more data than the predictions in *C. parapsilosis*. Of the studied long noncoding RNAs in *S. cerevisiae*, only two have experimentally validated functional association. One of them was predicted in our reduced dataset of 55. Two further noncoding transcripts [noncoding RNAs] were observed in several individual transcriptome assemblies, but did not meet the conservation threshold. The GO association of the transcripts in the analysis was broad, and is not yet sufficient to classify the precise functions. Since only one of the transcripts in the 55 available has a proven function, the stochasticity of the analysis would not allow a validation of the pipeline. There are Haploid strains of cerevisiae available, making deletion mutants more straightforward. Additionally, the pipeline can be used to extensively predict more features in different species. Given sufficient amount of new RNA sequencing data. The project

has many more pathways to explore from its current state. Only one of the transcripts has been studied in more detail. MAD1 / MADCaT and its impact on the yeast were unexpected, and the scope of the analysis expanded as the project went on. Amongst several interesting traits, the impact on temperature sensitivity is likely to be the most important one. Our estimation of the functionality concerns the composition of the cell wall, with regulatory elements involved in cell wall composition and efflux pump regulation. The cell wall composition changes the interaction of the yeast with its surroundings. This potentially important discovery has to be explored in greater detail, in order to investigate the transcript as a potential drug target. Apart from the clinical impact, the mechanisms by themselves are intriguing, and not yet completely clear. An interesting feature is the seeming ability of the yeast to switch from a coding reverse strand to a noncoding forward strand under different conditions. Overall the amount of phenotypes observed was surprisingly high. Other knockout studies, even with guided RNA sequencing led us to initially expect no phenotypes. Ultimately, we observed four out of five observable phenotypes our long noncoding annotation pipeline. The project carries much potential for downstream evaluations on a variety of ways. Apart from additional deletion mutants, the implemented pipelines can help the detection and evaluation of noncoding features in other organisms.

6 Conclusions

- A new pipeline for the exploratory detection and annotation of noncoding RNA in nonmodel species of yeast has been developed. We applied the detection to the pathogenic yeast *Candida parapsilosis* and *Saccharomyces cerevisiae* validating transcripts by deletion and phenotype screening, with a high rate of success in phenotype studies.

- Potential noncoding transcripts in the pathogenic yeast *Candida parapsilosis* were investigated. We concluded that sequence conservation and secondary structure formation broadly behave like noncoding transcripts in higher eukaryotes, and the expression profile suggests functionality.

- Certain noncoding transcribed RNAs might be an important mediator for temperature resistance in the yeast *candida parapsilosis*. One of the investigated transcripts (MAD) is predicted to form similar secondary structures in forward and reverse strand. Its function, potentially in combination with an ORF on the reverse strand, enables the yeast to withstand physiological temperatures. This makes the transcript a potentially important factor in the yeasts virulence.

- By developing a specific software solution, we improved the accuracy of the analysis of Allele specific expression in nonmodel species lacking a phased genome. Our approach increases the precision of analysis in heterogeneous populations.

- The cancer derived cell line THP-1 is capable of detecting and responding differently to various pathogens. Its specific transcription profiles can inform on specific defense strategies towards distinct pathogens.

- Although THP-1 can be used to investigate the impact of pathogens, time course analysis do not seem to introduce a uniform response in this context, and may therefore require a more sophisticated experimental setup.

References

Agostini, F., Zanzoni, A., Klus, P., Marchese, D., Cirillo, D., and Tartaglia, G. G. (2013). CatRAPID omics: A web server for large-scale prediction of protein-RNA interactions. *Bioinformatics*, 29(22):2928–2930.

Aickin, M. and Gensler, H. (1996). Adjusting for multiple testing when reporting research results: The Bonferroni vs Holm methods. *American Journal of Public Health*, 86(5):726–728.

Amorim-Vaz, S., Delarze, E., Ischer, F., Sanglard, D., and Coste, A. T. (2015). Examining the virulence of Candida albicans transcription factor mutants using Galleria mellonella and mouse infection models. *Frontiers in Microbiology*, 6(MAY):1–14.

Anders, S., Huber, W., Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M., Mortazavi, A., Williams, B., McCue, K., Schaeffer, L., Wold, B., Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O., He, A., Marra, M., Snyder, M., Jones, S., Licatalosi, D., Mele, A., Fak, J., Ule, J., Kayikci, M., Chi, S., Clark, T., Schweitzer, A., Blume, J., Wang, X., Darnell, J., Darnell, R., Smith, A., Heisler, L., Mellor, J., Kaper, F., Thompson, M., Chee, M., Roth, F., Giaever, G., Nislow, C., Marioni, J., Mason, C., Mane, S., Stephens, M., Gilad, Y., Wang, L., Feng, Z., Wang, X., Wang, X., Zhang, X., Robinson, M., Smyth, G., Whitaker, L., Robinson, M., McCarthy, D., Smyth, G., Robinson, M., Smyth, G., Cameron, A., Trivedi, P., Robinson, M., Oshlack, A., Loader, C., McCullagh, P., Nelder, J., Agresti, A., Engström, P., Tommei, D., Stricker, S., Smith, A., Pollard, S., Bertone, P., Morrissy, A., Morin, R., Delaney, A., Zeng, T., McDonald, H., Jones, S., Zhao, Y., Hirst, M., Marra, M., Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S., Habegger, L., Rozowsky, J., Shi, M., Urban, A., Hong, M., Karczewski, K., Huber, W., Weissman, S., Gerstein, M., Korbel, J., Snyder, M., Benjamini, Y., Hochberg, Y., Bullard, J., Purdom, E., Hansen, K., Dudoit, S., Bloom, J., Khan, Z., Kruglyak, L., Singh, M., Caudy, A., Smyth, G., Smyth, G., Lönnstedt, I., Speed, T., Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J., Zhang, J., Bliss, C., Fisher, R., Clark, S., Perry, J., Lawless, J., Saha, K., Paul, S., Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169.

Ariel, F., Romero-Barrios, N., Jgu, T., Benhamed, M., and Crespi, M. (2015). Battles and hijacks: noncoding transcription in plants. *Trends Plant Sci.*

Arnaud, M. B., Costanzo, M. C., Skrzypek, M. S., Binkley, G., Lane, C., Miyasato, S. R., and Sherlock, G. (2005). The Candida Genome Database (CGD), a community resource for Candida albicans gene and protein information. *Nucleic Acids Research*, 33(DATABASE ISS.):358–363.

Badrane, H., Cheng, S., Nguyen, M. H., Jia, H. Y., Zhang, Z., Weisner, N., and Clancy, C. J. (2005). Candida albicans IRS4 contributes to hypal formation and virulence after the initial stages of disseminated candidiasis. *Microbiology*, 151(9):2923–2931.

Banerjee Chitnis, A., Jadhav, S., Bhawalkar, J., and Chaudhury, S. (2011). Hypothesis testing, type I and type II errors. *IPJ.*

Barendsen, N., Mueller, M., and Chen, B. (2008). Inhibition of TPA-induced monocytic differentiation in THP-1 human monocytic leukemic cells by staurosporine, a potent protein kinase C inhibitor. *Leuk Res.*

Belardelli, F. and Gresser, I. (2010). The neglected role of type I interferon in the T-cell response: implications for its clinical use. *Immunology Today*.

Bell, G. D. M., Kane, N. C., Rieseberg, L. H., and Adams, K. L. (2013). RNA-seq analysis of allele-specific expression, hybrid effects, and regulatory divergence in hybrids compared with their parents from natural populations. *Genome Biology and Evolution*, 5(7):1309–1323.

Bennett, R. J. (2015). The parasexual lifestyle of Candida albicans. *Current Opinion in Microbiology*, 28:10–17.

Berman, J. and Sudbery, P. E. (2002). Candida albicans: A molecular revolution built on lessons from budding yeast. *Nature Reviews Genetics*, 3(12):918–932.

Bertone, P., Stolc, V., Royce, T., Rozowsky, J., Urban, A., Zhu, X., Rinn, J., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M., and Snyder, M. (2004). Global identification of human transcribed sequences with genome tiling arrays. *Science*.

Bishop, A., Lane, R., Beniston, R., Chapa-y Lazo, B., Smythe, C., and Sudbery, P. (2010). Hyphal growth in Candida albicans requires the phosphorylation of Sec2 by the Cdc28-Ccn1/Hgc1 kinase. *The EMBO journal*, 29(17):2930–42.

Bolard, J., Legrand, P., Heitz, F., and Cybulska, B. (1991). One-sided action of amphotericin B on cholesterol-containing membranes is determined by its self-association in the medium. *Biochemistry*.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

Bonfietti, L. X., Martins, M. d. A., Szeszs, M. W., Pukiskas, S. B. S., Purisco, S. U., Pimentel, F. C., Pereira, G. H., Silva, D. C., Oliveira, L., and Melhem, M. d. S. C. (2012). Prevalence, distribution and antifungal susceptibility profiles of Candida parapsilosis, Candida orthopsilosis and Candida metapsilosis bloodstream isolates. *Journal of Medical Microbiology*, 61(PART7):1003–1008.

Boube, M., Joulia, L., Cribbs, D. L., and Bourbon, H. M. (2002). Evidence for a MED of RNA polymerase II transcriptional regulation conserved from yeast to man. *Cell*, 110:143–151.

Branchini, M. L., Pfaller, M. A., Rhine-Chalberg, J., Frempong, T., and Isenberg, H. D. (1994). Genotypic variation and slime production among blood and catheter isolates of Candida parapsilosis. *Journal of Clinical Microbiology*, 32(2):452–456.

Brown, A. J. P., Brown, G. D., Netea, M. G., and Gow, N. A. R. (2014). Metabolism impacts upon candida immunogenicity and pathogenicity at multiple levels. *Trends in Microbiology*, 22(11):614–622.

Bryan, A. C., Rodeheffer, M. S., Wearn, C. M., and Shadel, G. S. (2002). Sls1p is a membrane-bound regulator of transcription-coupled processes involved in Saccharomyces cerevisiae mitochondrial gene expression. *Genetics*, 160(1):75–82.

Bumgarner, R. (2013). DNA microarrays: Types, Applications and their future. *Curr Protoc Mol Biol.*, 6137(206):1–17.

Cai, B., Song, X., Cai, J., and Zhang, S. (2014). HOTAIR: a cancer-related long non-coding RNA. *Neoplasma*.

Carmi, I. (2006). Molecular Biology Select. *Cell*.

Carninci, P., Kasukawa, S., Katayama, J., Gough, C., Frith, N., Maeda, R., Oyama, T., Ravasi, B., Lenhard, C., Wells, R., Kodzius, K., Shimokawa, V., Bajic, S., E. Brenner, S., Batalov, A. R., Forrest, R., and Zavolan, M. (2006). The Transcriptional Landscape of the Mammalian Genome. *Science*, 309(5740):1559–1563.

Casadevall, A. and Pirofski, L. (1999). Host-pathogen interactions: redefi ning the basic concepts of virulence and pathogenicity. *Infection and immunity*, 67(8):3703–3713.

Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., and Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome Biology*, 16(1):195.

Chandra, J., Kuhn, D. M., Mukherjee, P. K., Hoyer, L. L., McCormick, T., Ghannoum, M. A., Mahmoud, a., Cormick, T. M. C., Ghannoum, M. A., Mitchell, J. S. F., and P., A. (2001). Genetic Control of Candida Albicans Biofilm Development. *National Reviiew of Microbiology*, 9(18):109–118.

Chauvel, M., Nesseir, A., Cabral, V., Znaidi, S., Goyard, S., Bachellier-Bassi, S., Firon, A., Legrand, M., Diogo, D., Naulleau, C., Rossignol, T., and D'Enfert, C. (2012). A Versatile Overexpression Strategy in the Pathogenic Yeast Candida albicans: Identification of Regulators of Morphogenesis and Fitness. *PLoS ONE*, 7(9).

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Karra, K., Krieger, C. J., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., Simison, M., Weng, S., and Wong, E. D. (2012). Saccharomyces Genome Database: The genomics resource of budding yeast. *Nucleic Acids Research*, 40(D1):700–705.

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

Citiulo, F., Moran, G. P., Coleman, D. C., and Sullivan, D. J. (2009). Purification and germination of Candida albicans and Candida dubliniensis chlamydospores cultured in liquid media. *FEMS Yeast Research*, 9(7):1051–1060.

Colombo, A. L. and Guimarães, T. (2003). Epidemiology of hematogenous infections due to Candida spp. *Revista da Sociedade Brasileira de Medicina Tropical*, 36(5):599–607.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., and Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1):13.

Conti, H. R., Huppler, A. R., Whibley, N., and Gaffen, S. L. (2014). Animal Models for Candidiasis. *Curr Protoc Immunol.*

Cotter, G., Doyle, S., and Kavanagh, K. (2000). Development of an insect model for the in vivo pathogenicity testing of yeasts. *FEMS Immunology and Medical Microbiology*, 27(2):163–169.

Crampin, H. (2005). Candida albicans hyphae have a Spitzenkorper that is distinct from the polarisome found in yeast and pseudohyphae. *Journal of Cell Science*, 118(13):2935–2947.

Degner, J. F., Marioni, J. C., Pai, A. A., Pickrell, J. K., Nkadori, E., Gilad, Y., and Pritchard, J. K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–3212.

Delaloye, J. and Calandra, T. (2014). Invasive candidiasis as a cause of sepsis in the critically ill patient. *Virulence*, 5(1):161–169.

Denning, D. (2003). Echinocandin antifungal drugs. *Lancet.*

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J., Lipovich, L., Gonzalez, J., Thomas, M., Davis, C., Shiekhattar, R.,

Gingeras, T., Hubbard, T., Notredame, C., Harrow, J., and Guig, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*.

Djebali S, L. J. (2012). Landscape of transcription in human cells. *Nature*, 489(7414):101–108.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

Dobin, A., Gingeras, T. R., and Spring, C. (2016). Optimizing RNA-Seq Mapping with STAR. *Methods Mol Biol*.

Donovan, P. D., Schroder, M. S., Higgins, D. G., and Butler, G. (2016). Identification of non-coding RNAs in the Candida parapsilosis species group. *PLoS ONE*, 11(9):1–14.

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

Dumitru, R., Navarathna, D. H. M. L. P., Semighini, C. P., Elowsky, C. G., Dumitru, R. V., Dignard, D., Whiteway, M., Atkin, A. L., and Nickerson, K. W. (2007). In vivo and in vitro anaerobic mating in Candida albicans. *Eukaryotic Cell*, 6(3):465–472.

Dwight, S. S., Balakrishnan, R., Christie, K. R., Costanzo, M. C., Dolinski, K., Engel, S. R., Feierbach, B., Fisk, D. G., Hirschman, J., Hong, E. L., Nash, R. S., Sethuraman, A., Starr, B., Theesfeld, C. L., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Weng, S., Botstein, D., and Cherry, J. M. (2004). Saccharomyces genome database: Underlying principles and organisation. *Brief Bioinform*, 5(1):9–22.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10(1):48.

ENCODEProjectConsortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816.

Ene, I. V., Brunke, S., Brown, A. J. P., and Hube, B. (2014). Metabolism in fungal pathogenesis. *Cold Spring Harbor Perspectives in Medicine*, 4(12):1–21.

Falagas, M., Betsi, G., and Athanasiou, S. (2006). Probiotics for prevention of recurrent vulvovaginal candidiasis: a review. *J Antimicrob Chemother.*

Falagas, M. E., Roussos, N., and Vardakas, K. Z. (2010). Relative frequency of albicans and the various non-albicans Candida spp among candidemia isolates from inpatients in various parts of the world: A systematic review. *International Journal of Infectious Diseases*, 14(11):e954–e966.

Fidel, P. L., Vazquez, J. A., and Sobel, J. D. (1999). Candida glabrata: review of epidemiology, pathogenesis, and clinical disease with comparison to C. albicans. *Clinical microbiology reviews*, 12(1):80–96.

Fink, D. (1997). A Compendium of Conjugate Priors. *Environmental Statistics Group Department of Biology Montana State Univeristy*.

Fuchs, B., O'Brien, E., Khoury, J., and Mylonakis, E. (2010). Methods for using Galleria mellonella as a model host to study fungal pathogenesis. *Virulence*.

Gabaldon, T., Naranjo-Ortiz, M. A., and Marcet-Houben, M. (2016). Evolutionary genomics of yeast pathogens in the Saccharomycotina. *FEMS Yeast Research*, 16(6):1–10.

Gácser, A., Trofa, D., Schäfer, W., and Nosanchuk, J. D. (2007). Targeted gene deletion in Candida parapsilosis demonstrates the role of secreted lipase in virulence. *Journal of Clinical Investigation*, 117(10):3049–3058.

Gerami-Nejad, M., Zacchi, L. F., McClellan, M., Matter, K., and Berman, J. (2013). Shuttle vectors for facile gap repair cloning and integration into a neutral locus in Candida albicans. *Microbiology (United Kingdom)*, 159(PART3):565–579.

Gow, N. A. R., van de Veerdonk, F. L., Brown, A. J. P., and Netea, M. G. (2011). Candida albicans morphogenesis and host defence: discriminating invasion from colonization. *Nature Reviews Microbiology*, 10(2):112–122.

Gregory, T. R. (2007). Coincidence, coevolution, or causation? DNA content, cellsize, and the C-value enigma. *Biological Reviews*, 76(1):65–101.

Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guig??, R., and Sammeth, M. (2012). Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Research*, 40(20):10073–10083.

Grózer, Z., Tóth, A., Tóth, R., Kecskeméti, A., Vágvölgyi, C., Nosanchuk, J. D., Szekeres, A., and Gácser, A. (2015). Candida parapsilosis produces prostaglandins from exogenous arachidonic acid and OLE2 is not required for their synthesis. *Virulence*, 6(1):85–92.

Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R., and Hofacker, I. L. (2008). The Vienna RNA websuite. *Nucleic acids research*, 36(Web Server issue):70–74.

Guida, A., Lindstadt, C., Maguire, S. L., Ding, C., Higgins, D. G., Corton, N. J., Berriman, M., and Butler, G. (2011). Using RNA-seq to determine the transcriptional landscape and the hypoxic response of the pathogenic yeast Candida parapsilosis. *BMC Genomics*, 12(1):628.

Guinea, J. (2014). Global trends in the distribution of Candida species causing candidemia. *Clinical Microbiology and Infection*, 20(6):5–10.

Gutschner, T., Hammerle, M., Eissmann, M., Hsu, J., Kim, Y., Hung, G., Revenko, A., Arun, G., Stentrup, M., Gross, M., Zrnig, M., MacLeod, A., Spector, D., and Diederichs, S. (2013). The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res.*

Hagen, J. B. (2000). The origins of bioinformatics. *Nature Reviews Genetics*, 1(3):231–236.

Hanners, N. W., Eitson, J. L., Usui, N., Richardson, R. B., Wexler, E. M., Konopka, G., and Schoggins, J. W. (2016). Western Zika Virus in Human Fetal Neural Progenitors Persists Long Term with Partial Cytopathic and Limited Immunogenic Effects. *Cell Reports*, 15(11):2315–2322.

Harrington, C., Lin, E., Olson, M., and Eshleman, J. (2013). Fundamentals of pyrosequencing. *Arch Pathol Lab Med.*

Hawser, S. P. and Douglas, L. J. (1994). Biofilm formation by Candida species on the surface of catheter materials in vitro. *Infection and Immunity*, 62(3):915–921.

Hebecker, B., Vlaic, S., Conrad, T., Bauer, M., Brunke, S., Kapitan, M., Linde, J., Hube, B., and Jacobsen, I. D. (2016). Dual-species transcriptional profiling during systemic candidiasis reveals organ-specific host-pathogen interactions. *Scientific Reports*, 6(1):36055.

Hirano, R., Sakamoto, Y., Kudo, K., and Ohnishi, M. (2015). Retrospective analysis of mortality and Candida isolates of 75 patients with candidemia: A single hospital experience. *Infection and Drug Resistance*, 8:199–205.

Hoffmann, C., Dollive, S., Grunberg, S., Chen, J., Li, H., Wu, G. D., Lewis, J. D., and Bushman, F. D. (2013). Archaea and Fungi of the Human Gut Microbiome: Correlations with Diet and Bacterial Residents. *PLoS ONE*, 8(6).

Holland, L. M., Schröder, M. S., Turner, S. A., Taff, H., Andes, D., Grózer, Z., Gácser, A., Ames, L., Haynes, K., Higgins, D. G., and Butler, G. (2014). Comparative Phenotypic Analysis of the Major Fungal Pathogens Candida parapsilosis and Candida albicans. *PLoS Pathogens*, 10(9).

Hood, L. and Rowen, L. (2013). The human genome project: big science transforms biology and medicine. *Genome Medicine*, 5(9):79.

Houseley, J., Rubbi, L., Grunstein, M., Tollervey, D., and Vogelauer, M. (2008). A ncRNA Modulates Histone Modification and mRNA Induction in the Yeast GAL Gene Cluster. *Molecular Cell*, 32(5):685–695.

HumanGenomeProjectConsortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

Huynh, C., Sacks, D. L., and Andrews, N. W. (2006). A Leishmania amazonensis ZIP family iron transporter is essential for parasite replication within macrophage phagolysosomes. *The Journal of Experimental Medicine*, 203(10):2363–2375.

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T. R., Presner, J. R., Evans, J. R., Zhao, S., Poliakov, A., Cao, X., Dhanasekaran, S. M., Wu, Y.-M., Robinson, D. R., Beer, D. G., Feng, F., Iyer, H. K., and Chinnnaiyan, A. M. (2015). The Landscape of Long Noncoding RNAs in the Human Transcriptome. *Nat Genet.*, 47(3):199–208.

Jacobsen, I. D. (2014). Galleria mellonella as a model host to study virulence of Candida. *Virulence*, 5(2):237–239.

Jalali, S., Bhartiya, D., Lalwani, M. K., Sivasubbu, S., and Scaria, V. (2013). Systematic Transcriptome Wide Analysis of lncRNA-miRNA Interactions. *PLoS ONE*, 8(2).

Janbon, G., Ormerod, K. L., Paulet, D., Byrnes, E. J., Yadav, V., Chatterjee, G., Mullapudi, N., Hon, C. C., Billmyre, R. B., Brunel, F., Bahn, Y. S., Chen, W., Chen, Y., Chow, E. W. L., Coppée, J. Y., Floyd-Averette, A., Gaillardin, C., Gerik, K. J., Goldberg, J., Gonzalez-Hilarion, S., Gujja, S., Hamlin, J. L., Hsueh, Y. P., Ianiri, G., Jones, S., Kodira, C. D., Kozubowski, L., Lam, W., Marra, M., Mesner, L. D., Mieczkowski, P. a., Moyrand, F., Nielsen, K., Proux, C., Rossignol, T., Schein, J. E., Sun, S., Wollschlaeger, C., Wood, I. a., Zeng, Q., Neuvéglise, C., Newlon, C. S., Perfect, J. R., Lodge, J. K., Idnurm, A., Stajich, J. E., Kronstad, J. W., Sanyal, K., Heitman, J., Fraser, J. a., Cuomo, C. a., and Dietrich, F. S. (2014). Analysis of the Genome and Transcriptome of Cryptococcus neoformans var. grubii Reveals Complex RNA Expression and Microevolution Leading to Virulence Attenuation. *PLoS Genetics*, 10(4).

Jiang, Y., Zhang, N. R., and Li, M. (2017). SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. *Genome Biology*, 18(1):74.

Johnsson, P., Lipovich, L., Grandér, D., and Morris, K. V. (2014). Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochimica et biophysica acta*, 1840(3):1063–71.

Kaur, G. and Dufour, J. M. (2012). Cell lines: Valuable tools or useless artifacts. *Spermatogenesis*, 2(1):1–5.

Kaur, R., Domergue, R., Zupancic, M. L., and Cormack, B. P. (2005). A yeast by any other name: Candida glabrata and its interaction with the host. *Current Opinion in Microbiology*, 8(4):378–384.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254.

Kharchenko, P. V., Silberstein, L., and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742.

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*.

Kircher, M. and Kelso, J. (2010). High-throughput DNA sequencing - Concepts and limitations. *BioEssays*, 32(6):524–536.

Knight, S. a. B., Vilaire, G., Lesuisse, E., and Dancis, A. (2005). Iron Acquisition from Transferrin by. *Society*, 73(9):5482–5492.

Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L., and Gao, G. (2007). CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*, 35(SUPPL.2):345–349.

Kuhn, D. M., Chandra, J., Mukherjee, P. K., and Ghannoum, M. a. (2002). Comparison of Bio lms Formed by. *Society*, 70(2):878–888.

Kuras, L., Barbey, R., and Thomas, D. (1997). Assembly of a bZIP-bHLH transcription activation complex: Formation of the yeast Cbf1-Met4-Met28 complex is regulated through Met28 stimulation of Cbf1 DNA binding. *EMBO Journal*, 16(9):2441–2451.

Kuras, L., Cherest, H., Surdin-Kerjan, Y., and Thomas, D. (1996). A heteromeric complex containing the centromere binding factor 1 and two basic leucine zipper factors, Met4 and Met28, mediates the transcription activation of yeast sulfur metabolism. *The EMBO journal*, 15(10):2519–2529.

Lachke, S. A., Joly, S., Daniels, K., and Soll, D. R. (2002). Phenotypic switching and filamentation in Candida glabrata. *Microbiology*, 148(9):2661–2674.

Lachke, S. A., Srikantha, T., Tsai, L. K., Daniels, K., and Soll, D. R. (2000). Phenotypic switching in Candida glabrata involves phase-specific regulation of the metallothionein gene MT-II and the newly discovered hemolysin gene HLP. *Infection and Immunity*, 68(2):884–895.

Lamping, E., Monk, B. C., Niimi, K., Holmes, A. R., Tsao, S., Tanabe, K., Niimi, M., Uehara, Y., and Cannon, R. D. (2007). Characterization of three classes of membrane proteins involved in fungal azole resistance by functional hyperexpression in Saccharomyces cerevisiae. *Eukaryotic Cell*, 6(7):1150–1165.

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559.

Langmead, B. and Salzberg, S. (2013). Fast gapped-read alignment with Bowtie2. *Nature methods*, 9(4):357–359.

Le, S., Josse, J., and Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysi. *Journal of Statistical Software*.

Lee, M. R., Sheng, W. H., Hung, C. C., Yu, C. J., Lee, L. N., and Hsueh, P. R. (2015). Mycobacterium abscessus complex infections in humans. *Emerging Infectious Diseases*, 21(9):1638–1646.

Leinonen, R., Sugawara, H., and Shumway, M. (2011). The sequence read archive. *Nucleic Acids Research*, 39(SUPPL. 1):2010–2012.

Leland, D. S. and Ginocchio, C. C. (2007). Role of cell culture for virus detection in the age of technology. *Clinical Microbiology Reviews*, 20(1):49–78.

LePage, D. F. and Conlon, R. A. (2007). Animal models for disease. *Methods in Molecular Medicine*, 129:1–18.

Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, P., Piao, Y., Shon, H. S., and Ryu, K. H. (2015). Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics*, 16(1):347.

Liang, W.-C., Fu, W.-m., Wong, C.-W., Wang, Y., Wang, W.-M., Hu, G.-X., Zhang, L., Xiao, L.-J., Wan, D. C.-C., Zhang, J.-F., and Waye, M. M.-Y. (2015). The lncRNA H19 promotes epithelial to mesenchymal transition by functioning as miRNA sponges in colorectal cancer. *Oncotarget*, 6(26):22513–25.

Linde, J., Duggan, S., Weber, M., Horn, F., Sieber, P., Hellwig, D., Riege, K., Marz, M., Martin, R., Guthke, R., and Kurzai, O. (2015). Defining the transcriptomic landscape of Candida glabrata by RNA-Seq. *Nucleic Acids Research*, 43(3):1392–1406.

Lindsay, A. K., Morales, D. K., Liu, Z., Grahl, N., Zhang, A., Willger, S. D., Myers, L. C., and Hogan, D. A. (2014). Analysis of Candida albicans Mutants Defective in the Cdk8 Module of Mediator Reveal Links between Metabolism and Biofilm Formation. *PLoS Genetics*, 10(10).

Linhares, I. M., Witkin, S. S., Miranda, S. D., Fonseca, A. M., Pinotti, J. A., and Ledger, W. J. (2001). Differentiation Between Women With Vulvovaginal Symptoms Who are Positive or Negative for Candida Species by Culture. *Infectious Diseases in Obstetrics and Gynecology*, 9(4):221–225.

Liu, T., Ren, X., Xiao, T., Yang, J., Xu, X., Dong, J., Sun, L., Chen, R., and Jin, Q. (2013). Identification and characterisation of non-coding small RNAs in the pathogenic filamentous fungus Trichophyton rubrum. *BMC genomics*, 14(1):931.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13):1675–1680.

Lockhart, S. R., Messer, S. A., Pfaller, M. A., and Diekema, D. J. (2008). Geographic distribution and antifungal susceptibility of the newly described species Candida orthopsilosis and Candida metapsilosis in comparison to the closely related species Candida parapsilosis. *Journal of Clinical Microbiology*, 46(8):2659–2664.

Lockhart, S. R., Pujol, C., Daniels, K. J., Miller, M. G., Johnson, A. D., Pfaller, M. A., and Soil, D. R. (2002). In Candida albicans, white-opaque switchers are homozygous for mating type. *Genetics*, 162(2):737–745.

Lohse, M. B. and Johnson, A. D. (2010). White-opaque switching in Candida albicans. *Curr Opin Microbiol*, 12(6):650–654.

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):550.

Mallory, A. and Shkumatava, A. (2016). Europe PMC Funders Group LncRNAs in Vertebrates : Advances and Challenges. *Biochimie*, pages 3–14.

Malone, J. H. and Oliver, B. (2011). Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biology*, 9(1):34.

Maretty, L., Sibbesen, J. A., Krogh, A., Heber, S., Alekseyev, M., Sze, S.-H., Tang, H., Pevzner, P., Trapnell, C., Williams, B., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M., Salzberg, S., Wold, B., Pachter, L., Li, W., Feng, J., Jiang, T., Li, J., Jiang, C.-R., Brown, J., Huang, H., Bickel, P., Li, W., Jiang, T., Mezlini, A., Smith, E., Fiume, M., Buske, O., Savich, G., Shah, S., Aparicion, S., Chiang, D., Goldenberg, A., Brudno, M., Behr, J., Kahles, A., Zhong, Y., Sreedharan, V., Drewe, P., Rätsch, G., Tomescu, A., Kuosmanen, A., Rizzi, R., Mäkinen, V., Veli, M., Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S., Djebali, S., Davis, C., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G., Khatun, J., Williams, B., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R., Alioto, T., Antoshechkin, I., Baer, M., Bar, N., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Grabherr, M., Haas, B., Yassour, M., Levin, J., Thompson, D., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., Meyer, L., Zweig, A., Hinrichs, A., Karolchik, D., Kuhn, R., Wong, M., Sloan, C., Rosenbloom, K., Roe, G., Rhead, B., Raney, B., Pohl, A., Malladi, V., Li, C., Lee, B., Learned, K., Kirkup,

V., Hsu, F., Heitner, S., Harte, R., Haeussler, M., Guruvadoo, L., Goldman, M., Giardine, B., Fujita, P., Dreszer, T., Diekhans, M., Cline, M., Clawson, H., Barber, G., Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Sammeth, M., Au, K., Sebastiano, V., Afshar, P., Durruthy, J., Lee, L., Williams, B., van Bakel, H., Schadt, E., Reijo-Pera, R., Underwood, J., Wong, W., Schulz, M., Zerbino, D., Vingron, M., Birney, E., Trapnell, C., Hendrickson, D., Sauvageau, M., Goff, L., Rinn, J., Pachter, L., Langmead, B., and Salzberg, S. (2014). Bayesian transcriptome assembly. *Genome Biology*, 15(10):501.

Markowetz, F. (2017). All biology is computational biology. *PLoS Biology*, 15(3):4–7.

Martinez, O., Johnson, J. C., Honko, A., Yen, B., Shabman, R. S., Hensley, L. E., Olinger, G. G., and Basler, C. F. (2013). Ebola Virus Exploits a Monocyte Differentiation Program To Promote Its Entry. *Journal of Virology*, 87(7):3801–3814.

Mattei, E., Pietrosanto, M., Ferr??, F., and Helmer-Citterich, M. (2015). Web-Beagle: A web server for the alignment of RNA secondary structures. *Nucleic Acids Research*, 43(W1):W493–W497.

McHugh, C. A., Chen, C.-K., Chow, A., Surka, C. F., Tran, C., McDonel, P., Pandya-Jones, A., Blanco, M., Burghard, C., Moradian, A., Sweredoski, M. J., Shishkin, A. A., Su, J., Lander, E. S., Hess, S., Plath, K., and Guttman, M. (2015). The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature*, 521(7551):232–236.

McKenna, A., Matthew, H., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*.

Millar, J. A., Valdés, R., Kacharia, F. R., Landfear, S. M., Cambronne, E. D., and Raghavan, R. (2015). Coxiella burnetii and Leishmania mexicana residing within similar parasitophorous vacuoles elicit disparate host responses. *Frontiers in microbiology*, 6(August):794.

Miyazaki, H., Miyazaki, Y., Geber, A., Parkinson, T., Hitchcock, C., Falconer, D. J., Ward, D. J., Marsden, K., and Bennett, J. E. (1998). Fluconazole resistance associated with drug efflux and increased transcription of a drug transporter gene, PDH1, in Candida glabrata. *Antimicrobial Agents and Chemotherapy*, 42(7):1695–1701.

Mortazavi, A., Williams, B., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*.

Moyes, D. L., Wilson, D., Richardson, J. P., Mogavero, S., Tang, S. X., Wernecke, J., Gratacap, R. L., Robbins, J., Runglall, M., Murciano, C., Blagojevic, M., Thavaraj, S., Hebecker, B., Kasper, L., Vizcay, G., Iancu, S. I., Kichik, N., Kurzai, O., Luo, T., Cota, E., Bader, O., Wheeler, R. T., Gutsmann, T., Hube, B., Naglik, J. R., Division, S. B., Sciences, B., and Immunology, G. M. (2016). Candidalysin is a fungal peptide toxin critical for mucosal infection. *Nature*, 532(7597):64–68.

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., Zhang, J., Weinstock, G. M., Isaacs, F., Rozowsky, J., and Gerstein, M. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biology*, 17(1):53.

Mukaka, M. M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3):69–71.

Murphy, A. and Kavanagh, K. (1999). Emergence of Saccharomyces cerevisiae as a human pathogen. *Enzyme and Microbial Technology*, 25(7):551–557.

Németh, T., Tóth, A., Szenzenstein, J., Horváth, P., Nosanchuk, J. D., Grózer, Z., Tóth, R., Papp, C., Hamari, Z., Vágvölgyi, C., and Gácser, A. (2013). Characterization of Virulence Properties in the C. parapsilosis Sensu Lato Species. *PLoS ONE*, 8(7):1–10.

Netea, M. G., Joosten, L. A. B., van der Meer, J. W. M., Kullberg, B.-J., and van de Veerdonk, F. L. (2015). Immune defence against Candida fungal infections. *Nature Reviews Immunology*, 15(10):630–642.

Neu, N., Malik, M., Lunding, A., Whittier, S., Alba, L., Kubin, C., and Saiman, L. (2009). Epidemiology of Candidemia at a Childrens Hospital, 2002 to 2006. *The Pediatric Infectious Disease Journal*, 28(9):806–809.

Nitsche, A. and Stadler, P. F. (2017). Evolutionary clues in lncRNAs. *Wiley Interdisciplinary Reviews: RNA*, 8(1):14–17.

Noble, W. S. (2010). NIH Public Access. *Nat Biotechnol.*, 27(12):1135–1137.

Nunn, M. A., Schäfer, S. M., Petrou, M. A., and Brown, J. R. M. (2007). Environmental source of Candida dubliniensis. *Emerging Infectious Diseases*, 13(5):747–750.

Nur, Y. (2014). Epidemiology and risk factors for invasive candidiasis. *Ther Clin Risk Manag.*

Nyirjesy, P., Alexander, A. B., and Weitz, M. V. (2005). Vaginal Candida parapsilosis: pathogen or bystander? *Infectious diseases in obstetrics and gynecology*, 13(1):37–41.

Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., Yamanaka, I., Kiyosawa, H., Yagi, K., Tomaru, Y., Hasegawa, Y., Nogami, A., Schönbach, C., Gojobori, T., Baldarelli, R., Hill, D. P., Bult, C., Hume, D. A., Quackenbush, J., Schriml, L. M., Kanapin, A., Matsuda, H., Batalov, S., Beisel, K. W., Blake, J. A., Bradt, D., Brusic, V., Chothia, C., Corbani, L. E., Cousins, S., Dalla, E., Dragani, T. A., Fletcher, C. F., Forrest, A., Frazer, K. S., Gaasterland, T., Gariboldi, M., Gissi, C., Godzik, A., Gough, J., Grimmond, S., Gustincich, S., Hirokawa, N., Jackson, I. J., Jarvis, E. D., Kanai, A., Kawaji, H., Kawasawa, Y., Kedzierski, R. M., King, B. L., Konagaya, A., Kurochkin, I. V., Lee, Y., Lenhard, B., Lyons, P. A., Maglott, D. R., Maltais, L., Marchionni, L., McKenzie, L., Miki, H., Nagashima, T., Numata, K., Okido, T., Pavan, W. J., Pertea, G., Pesole, G., Petrovsky, N., Pillai, R., Pontius, J. U., Qi, D., Ramachandran, S., Ravasi, T., Reed, J. C., Reed, D. J., Reid, J., Ring, B. Z., Ringwald, M., Sandelin, A., Schneider, C., Semple, C. A. M.,

Setou, M., Shimada, K., Sultana, R., Takenaka, Y., Taylor, M. S., Teasdale, R. D., Tomita, M., Verardo, R., Wagner, L., Wahlestedt, C., Wang, Y., Watanabe, Y., Wells, C., Wilming, L. G., Wynshaw-Boris, A., Yanagisawa, M., Yang, I., Yang, L., Yuan, Z., Zavolan, M., Zhu, Y., Zimmer, A., Carninci, P., Hayatsu, N., Hirozane-Kishikawa, T., Konno, H., Nakamura, M., Sakazume, N., Sato, K., Shiraki, T., Waki, K., Kawai, J., Aizawa, K., Arakawa, T., Fukuda, S., Hara, A., Hashizume, W., Imotani, K., Ishii, Y., Itoh, M., Kagawa, I., Miyazaki, A., Sakai, K., Sasaki, D., Shibata, K., Shinagawa, A., Yasunishi, A., Yoshino, M., Waterston, R., Lander, E. S., Rogers, J., Birney, E., and Hayashizaki, Y. (2002). Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, 420(6915):563–573.

Paixo, E. S., Barreto, F., Da Gloria Teixeira, M., Da Conceio N Costa, M., and Rodrigues, L. C. (2016). History, epidemiology, and clinical manifestations of Zika: A systematic review. *American Journal of Public Health*, 106(4):606–612.

Pammi, M., Holland, L., Butler, G., and Gacser, A. (2013). Candida parapsilosis is a Significant Neonatal Pathogen: A Systematic Review and Meta-Analysis. *Pediatr Infect Dis J*.

Pasko, M., SC, P., and AD., V. S. (1990). Fluconazole: a new triazole antifungal agent. *DICP.*

Patro, R., Mount, S. M., and Kingsford, C. (2013). Sailfish: Alignment-free Isoform Quantification from RNA-seq Reads using Lightweight Algorithms. *Nature Biotechnology*, 32(5):462–464.

Pontier, D. B. and Gribnau, J. (2011). Xist regulation and function eXplored. *Human Genetics*, 130(2):223–236.

Prabhakar, S., Qiao, Y., Hoshino, Y., Weiden, M., Canova, A., Giacomini, E., Coccia, E., and Pine, R. (2003). Inhibition of Response to Alpha Interferon by Mycobacterium tuberculosis. *Infect Immun*.

Pryszcz, L. P., Nemeth, T., Saus, E., Ksiezopolska, E., Hegedesova, E., Nosek, J., Wolfe, K. H., Gacser, A., and Gabaldon, T. (2015). The Genomic Aftermath of Hybridization in the Opportunistic Pathogen Candida metapsilosis. *PLoS Genetics*, 11(10):1–29.

Pujol, C., Daniels, K. J., Lockhart, S. R., Srikantha, T., Radke, J. B., Geiger, J., and Soll, D. R. (2004). The Closely Related Species Candida albicans and Candida dubliniensis Can Mate. *Society*, 3(4):1015–1027.

Raiffa, H. and Schlaifer, R. (1961). Applied Statistical Decision Theory, Division of Research. *Graduate School of Business Administration, Harvard*.

Reedy, J. L., Floyd, A. M., and Heitman, J. (2009). Mechanistic plasticity of sexual reproduction and meiosis in the Candida pathogenic species complex. *Curr Biol*.

Reyes, L., Davidson, M. K., Thomas, L. C., and Davis, J. K. (1999). Effects of Mycoplasma fermentans incognitus on differentiation of THP-1 cells. *Infection and Immunity*, 67(7):3188–3192.

Roetzer, A., Gabaldón, T., and Schüller, C. (2011). From Saccharomyces cerevisiae to Candida glabrata in a few easy steps: Important adaptations for an opportunistic pathogen. *FEMS Microbiology Letters*, 314(1):1–9.

Roilides, E., Farmaki, E., Evdoridou, J., Dotis, J., Hatziioannidis, E., Tsivitanidou, M., Bibashi, E., Filioti, I., Sofianou, D., Gil-Lamaignere, C., Mueller, F., and Kremenopoulos, G. (2004). Neonatal candidiasis: analysis of epidemiology, drug susceptibility, and molecular typing of causative isolates. *Eur J Clin Microbiol Infect Dis.*

Sabino, R., Sampaio, P., Rosado, L., Videira, Z., Grenouillet, F., and Pais, C. (2015). Probiotics for prevention of recurrent vulvovaginal candidiasis: a review. *Clin Microbiol Infect.*

Sai, S., Holland, L. M., Mcgee, C. F., Lynch, D. B., and Butler, G. (2011). Evolution of mating within the candida parapsilosis species group. *Eukaryotic Cell*, 10(4):578–587.

Sanger, F. and Coulson, A. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol.*

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467.

Santos MA, T. M. (1995). The CUG codon is decoded in vivo as serine and not leucine in Candida albicans. *Nucleic Acids Res*.

Saracli, M. A., Fothergill, A. W., Sutton, D. A., and Wiederhold, N. P. (2015). Detection of triazole resistance among Candida species by matrix-assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF MS). *Medical Mycology*, 53(7):736–742.

Sardi, J. C. O., Scorzoni, L., Bernardi, T., Fusco-Almeida, A. M., and Mendes Giannini, M. J. S. (2013). Candida species: Current epidemiology, pathogenicity, biofilm formation, natural antifungal products and new therapeutic options. *Journal of Medical Microbiology*, 62(PART1):10–24.

Sarkar, N. (1997). Polyadenylation of mRNA in prokaryotes. *Annu Rev Biochem*.

Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. K., Gerstein, M. B. M., Metzker, M., Mardis, E., Lieberman-Aiden, E., van Berkum, N., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B., Sabo, P., Dorschner, M., Sandstrom, R., Bernstein, B., Bender, M., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L., Lander, E., Dekker, J., Core, L., Waterfall, J., Lis, J., Licatalosi, D., Mele, A., Fak, J., Ule, J., Kayikci, M., Chi, S., Clark, T., Schweitzer, A., Blume, J., Wang, X., Darnell, J., Darnell, R., Bennett, S., Barnes, C., Cox, A., Davies, L., Brown, C., Service, R., Mardis, E., Wheeler, D., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C., Irzyk, G., Lupski, J., Chinault, C., Song, X.-z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D., Margulies, M., Weinstock, G., Gibbs, R., Rothberg, J., Walter, C., Kitzman, J., MacKenzie, A., Adey, A., Hiatt, J., Patwardhan,

R., Sudmant, P., Ng, S., Alkan, C., Qiu, R., Eichler, E., Shendure, J., Cokus, S., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C., Pradhan, S., Nelson, S., Pellegrini, M., Jacobsen, S., Down, T., Rakyan, V., Turner, D., Flicek, P., Li, H., Kulesha, E., Graf, S., Johnson, N., Herrero, J., Tomazou, E., Thorne, N., Backdahl, L., Herberth, M., Howe, K., Jackson, D., Miretti, M., Marioni, J., Birney, E., Hubbard, T., Durbin, R., Tavare, S., Beck, S., Smith, A., Greenbaum, D., Douglas, S., Long, M., Gerstein, M. B. M., Barrett, T., Troup, D., Wilhite, S., Ledoux, P., Evangelista, C., Kim, I., Tomashevsky, M., Marshall, K., Phillippy, K., Sherman, P., Muertter, R., Holko, M., Ayanbule, O., Yefanov, A., Soboleva, A., Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E., Holloway, E., Kurbatova, N., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rustici, G., Sharma, A., Williams, E., Adamusiak, T., Brandizi, M., Sklyar, N., Brazma, A., Stein, L., Fritz, M.-Y. H.-Y., Leinonen, R., Cochrane, G., Birney, E., Danecek, P., Auton, A., Abecasis, G., Albers, C., Banks, E., Depristo, M., Handsaker, R., Lunter, G., Marth, G., Sherry, S., McVean, G., Durbin, R., Mortazavi, A., Williams, B., McCue, K., Schaeffer, L., Wold, B., Habegger, L., Sboner, A., Gianoulis, T., Rozowsky, J., Agarwal, A., Snyder, M., Gerstein, M. B. M., Schatz, M., Langmead, B., Salzberg, S., Langmead, B., Schatz, M., Lin, J., Pop, M., Salzberg, S., Goecks, J., Nekrutenko, A., Taylor, J., Team, T. G., Afgan, E., Baker, D., Coraor, N., Chapman, B., Nekrutenko, A., Taylor, J., Langmead, B., Hansen, K., Leek, J., Green, R., Krause, J., Briggs, A., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.-Y. H.-Y., Hansen, N., Durand, E., Malaspinas, A.-S., Jensen, J., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H., Good, J., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E., Russ, C., Wang, E., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S., Schroth, G., Burge, C., Sultan, M., Schulz, M., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H., Yaspo, M.-L., Maher, C., Palanisamy, N., Brenner, J., Cao, X., Kalyana-Sundaram, S., Luo, S., Khrebtukova, I., Barrette, T., Grasso, C., Yu, J., Lonigro, R., Schroth, G., Kumar-Sinha, C., Chinnaiyan, A., Pflueger, D., Terry, S., Sboner, A., Habegger, L., Esgueva, R., Lin, P.-C., Svensson, M., Kitabayashi, N., Moss, B., MacDonald, T., Cao, X., Barrette, T., Tewari, A., Chee, M., Chinnaiyan, A., Rickman, D., Demichelis, F., Gerstein, M. B. M., Rubin, M., Wulff, B.-E., Sakurai, M., Nishikura, K., Pickrell, J., Marioni, J., Pai, A., Degner, J., Engelhardt, B., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., Pritchard, J., Montgomery, S., Sammeth, M., Gutierrez-Arcelus, M., Lach, R., Ingle, C., Nisbett, J., Guigo, R., Dermitzakis, E., and Mardis, E. (2011). The real cost of sequencing: higher than you think! *Genome biology*, 12(8):125.

Schmitz, S. U., Grote, P., and Herrmann, B. G. (2016). Mechanisms of long noncoding RNA function in development and disease. *Cellular and Molecular Life Sciences*, 73(13):2491–2509.

Schroeder, S. C., Zorio, D. A., Schwer, B., Shuman, S., and Bentley, D. (2004). A Function of Yeast mRNA Cap Methyltransferase, Abd1, in Transcription by RNA Polymerase II. *Molecular Cell*, 13(3):377–387.

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., and Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *Rna*, 22(6):839–851.

Segal, B. H. (2007). Role of macrophages in host defense against aspergillosis and strategies for immune augmentation. *The oncologist*, 12 Suppl 2(Supplement 2):7–13.

Sellam, A., Hogues, H., Askew, C., Tebbji, F., van Het Hoog, M., Lavoie, H., Kumamoto, C. a., Whiteway, M., and Nantel, A. (2010). Experimental annotation of the human pathogen Candida albicans coding and noncoding transcribed regions using high-resolution tiling arrays. *Genome biology*, 11(7):R71.

Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2015). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*

Sigurdsson, M. I., Saddic, L., Heydarpour, M., Chang, T.-W., Shekar, P., Aranki, S., Couper, G. S., Shernan, S. K., Seidman, J. G., Body, S. C., and Muehlschlegel, J. D. (2016). Allele-specific expression in the human heart and its application to postoperative atrial fibrillation and myocardial ischemia. *Genome Medicine*, 8(1):127.

Singh, R. and Parija, S. C. (2012). Candida parapsilosis: An emerging fungal pathogen. *Indian Journal of Medical Research*, 136(4):671–673.

Smilinich, N. J., Day, C. D., Fitzpatrick, G. V., Caldwell, G. M., Lossie, A. C., Cooper, P. R., Smallwood, A. C., Joyce, J. A., Schofield, P. N., Reik, W., Nicholls, R. D., Weksberg, R., Driscoll, D. J., Maher, E. R., Shows, T. B., and Higgins, M. J. (1999). A maternally methylated CpG island in KvLQT1 is associated with an antisense paternal transcript and loss of imprinting in Beckwith-Wiedemann syndrome. *Proceedings of the National Academy of Sciences*, 96(14):8064–8069.

Springer, N. M. and Stupar, R. M. (2007). Allele-Specific Expression Patterns Reveal Biases and Embryo-Specific Parent-of-Origin Effects in Hybrid Maize. *the Plant Cell Online*, 19(8):2391–2402.

Stajich, J., Berbee, M. L., Blackwell, M., Hibbett, D. S., James, T. Y., Spatafora, J. W., and Taylor, J. W. (2009). The Fungi. *Current Biology*, 19(18):R840–R845.

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., and Robinson, G. E. (2015). Big data: Astronomical or genomical? *PLoS Biology*, 13(7):1–11.

Sullivan, D. J., Moran, G. P., and Coleman, D. C. (2005). Candida dubliniensis: Ten years on. *FEMS Microbiology Letters*, 253(1):9–17.

Szabo, P., Hubner, K., Scholer, H., and Mann, J. (2002). Allele-specific expression of imprinted genes in mouse migratory primordial germ cells. *Mech Dev.*

Szabo, P. E. and Mann, J. R. (1995). Allele-specific expression and total expression levels of imprinted genes during early mouse development: implications for imprinting mechanism. *Genes Dev.*, 9:3097–3108.

Szilagyi, J., Foldi, R., Gesztelyi, R., Bayegan, S., Kardos, G., Juhasz, B., and Majoros, L. (2012). Comparison of the kidney fungal burden in experimental disseminated candidiasis by species of the Candida parapsilosis complex treated with fluconazole, amphotericin B and caspofungin in a temporarily neutropenic murine model. *Chemotherapy*.

Tóth, A., Cabral, V., Thuer, E., Bohner, F., Nmeth, T., Papp, C., Nimrichter, L., Molnar, G., Vagvolgyi, C., Gabaldon, T., Nosanchuk, J. D., and Gacser, A. (2017). Investigation of Candida parapsilosis virulence regulatory networks following host-pathogen interaction. *PLOS Pathogen*.

Tóth, R., Tóth, A., Papp, C., Jankovics, F., Vágvölgyi, C., Alonso, M. F., Bain, J. M., Erwig, L. P., and Gácser, A. (2014). Kinetic studies of Candida parapsilosis phagocytosis by macrophages and detection of intracellular survival mechanisms. *Frontiers in Microbiology*, 5(NOV):1–12.

Trapnell, C., Williams, B. a., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2011). Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nature Biotechnology*, 28(5):511–515.

Tsuchiya, S., Yamabe, M., Yamaguchi, Y., Kobayashi, Y., Konno, T., and Tada, K. (1980). Establishment and characterization of a human acute monocytic leukemia cell line (THP-1). *Int J Cancer*.

Turner, S. A. and Butler, G. (2014). The Candida pathogenic species complex. *Cold Spring Harbor Perspectives in Medicine*, 4(9):1–18.

Ucsc, T. and Browser, G. (2003). The UCSC Genome Browser. *Curr Protoc Bioinformatics*, pages 1–23.

Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H., and Bartel, D. P. (2012). Conserved Function of lincRNAs in Vertebrate Embryonic Despite Rapid Sequence Evolution. *Cell*, 147(7):1537–1550.

UniProtConsortium (2015). UniProt: A hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212.

Vilain, C., Libert, F., Venet, D., Costagliola, S., and Vassart, G. (2003). Small amplified RNA-SAGE: an alternative approach to study transcriptome from limiting amount of mRNA. *Nucleic Acids Research*, 31(6):e24–.

Vyas, V. K., Barrasa, M. I., and Fink, G. R. (2015). A Candida albicans CRISPR system permits genetic engineering of essential genes and gene families. *Science advances*, 1(3):e1500248.

Wadi, L., Meyer, M., Weiser, J., Stein, L. D., and Reimand, J. (2016). Impact of outdated gene annotations on pathway enrichment analysis. *Nature Methods*, 13(9):705–706.

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J. P., and Li, W. (2013). CPAT: Coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Research*, 41(6):1–7.

Werven, F. J. V., Neuert, G., Hendrick, N., Lardenois, A., Oudenaarden, A. V., Primig, M., and Amon, A. (2013). Transcription of two long noncoding RNAs mediates mating-type control of gametogenesis in budding yeast. *Cell*, 150(6):1170–1181.

Wikhe, K., Westermeyer, C., and Macreadie, I. G. (2007). Biological consequences of statins in Candida species and possible implications for human health. *Biochem Soc Trans*, 35(Pt 6):1529–32.

Wilusz, J. E., Sunwoo, H., and Spector, D. L. (2009). Long noncoding RNAs : functional surprises from the RNA world Long noncoding RNAs : functional surprises from the RNA world. *Genes Dev*, pages 1494–1504.

Wisplinghoff, H., Bischoff, T., Tallent, S. M., Seifert, H., Wenzel, R. P., and Edmond, M. B. (2004). Nosocomial Bloodstream Infections in US Hospitals: Analysis of 24,179 Cases from a Prospective Nationwide Surveillance Study. *Clinical Infectious Diseases*, 39(3):309–317.

Yamashita, A., Shichino, Y., and Yamamoto, M. (2016). The long non-coding RNA world in yeasts. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1859(1):147–154.

Yang, J. and Zhang, J. (2015). Human long noncoding RNAs are substantially less folded than messenger RNAs. *Mol Biol Evol*.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–1591.

Yano, J. and Fidel, Jr., P. L. (2011). Protocols for Vaginal Inoculation and Sample Collection in the Experimental Mouse Model of Candida vaginitis. *Journal of Visualized Experiments*, pages 1–7.

Zakharova, E., Grandhi, J., Wewers, M. D., and Gavrilin, M. A. (2010). Mycoplasma suppression of THP-1 cell TLR responses is corrected with antibiotics. *PLoS ONE*, 5(3):3–6.

Zervos, M. and Meunier, F. (1993). Fluconazole (Diflucan): a review. *Int J Antimicrob Agents.*

Zhao, R., Daniels, K. J., Lockhart, S. R., Yeater, K. M., Hoyer, L. L., and Soll, D. R. (2005). Unique aspects of gene expression during Candida albicans mating and possible G1 dependency. *Eukaryotic Cell*, 4(7):1175–1190.

Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE*, 9(1).