UC Merced

UC Merced Electronic Theses and Dissertations

Title

Evolution of the gene regulatory network controlling biofilm formation in Candida species

Permalink

https://escholarship.org/uc/item/43d0n661

Author

Gunasekaran, Deepika

Publication Date

2023

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-ShareAlike License, available at https://creativecommons.org/licenses/by-nc-sa/4.0/

Peer reviewed|Thesis/dissertation

eScholarship.org

Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, MERCED

Evolution of the gene regulatory network controlling biofilm formation in *Candida* species

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Quantitative and Systems Biology

by

Deepika Gunasekaran

Committee in charge:

Professor Gordon Bennett, Chair
Professor Aaron Hernday
Professor Suzanne Sindi
Professor Clarissa Nobile, Advisor
Professor David Ardell, Co-advisor

2023

Evolution of the gene regulatory network controlling biofilm formation in *Candida* species

Copyright
by
Deepika Gunasekaran, 2023
All rights reserved.

The dissertation of Deepika Gunasekaran, titled "Evolution of the gene regulatory network controlling biofilm formation in *Candida* species", is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Date

Professor Gordon Bennett, Chair

Date

Professor Aaron Hernday

Date

Professor Suzanne Sindi

Date

Professor Clarissa Nobile, Advisor

Date

Professor David Ardell, Co-advisor

University of California, Merced
2023

To

My parents and my husband.
To the family I was blessed with and the one I found.

Contents

List of Figures

List of Tables

Abbreviations

ALLR Average Log-likelihood Ratio.

BP Biological Process.

CC Cellular Component.

CDF Cumulative Distribution Function.

CGD Candida Genome Database.

CNV Copy Number Variation.

DLC Duplication Loss Coalescence.

FDR False Discovery Rate.

GATK The Genome Analysis Toolkit.

GO Gene Ontology.

GRIDSS Genome Rearrangement IDentification Software Suite.

GRN Gene Regulatory Network.

GSEA Gene Set Enrichment Analysis.

GTR General Time Reversible.

IC Information Content.

IDR Intrinsically Disordered Region.

IP Immuno-precipitated.

iTol The Interactive Tree Of Life.

JSD Jensen-Shannon Divergence.

KLD Kullback-Liebler Divergence.

MCL Markov Clustering.

MF Molecular Function.

MLST Multilocus Sequence Typing.

MSA Multiple Sequence Alignment.

MYA Million Years Ago.

NCBI National Center for Biotechnology Information.

PMF Probability Mass Function.

PPI Protein-Protein Interaction.

PPM Position Probability Matrix.

SNP Single Nucleotide Polymorphism.

SRA Sequence Read Archive.

STAG Species Tree from All Genes.

SV Structural Variation.

TF Transcription Factor.

TFBS Transcription Factor Binding Site.

VEP Variant Effect Predictor.

WGD Whole Genome Duplication.

Acknowledgments

It is my privilege to thank all the remarkable people who have guided me along my journey as a PhD student. I would like to start by expressing my deepest and sincere gratitude to my advisors and mentors, Prof. Clarissa Nobile and Prof. Dave Ardell. I would like to thank Prof. Nobile for her guidance and support throughout the course of my graduate studies. Thank you for your encouragement to pursue opportunities and collaborative projects that helped me grow as a scientist. I always admired your temperament and will try to emulate it throughout my career. I would also like to thank Prof. Ardell for his support and guidance. Thank you for helping me broaden my scientific ground and learn from multiple disciplines. I continue to learn so much from the both of you and am always grateful to you for helping me find my academic identity and voice. Your mentorship and encouragement have been pivotal in shaping both my academic and personal growth. I am beyond fortunate to have you as my advisors and I look forward to future work and collaborations together.

I am profoundly grateful for the unwavering support and encouragement I received from my amazing thesis committee, Prof. Gordon Bennett, Prof. Aaron Hernday and Prof. Suzanne Sindi. I had the opportunity to work closely and learn from each of you since my first year at graduate school. Prof. Bennett, you were one of the first professors I interviewed with and did a lab rotation with at UC Merced. You gave me the opportunity to drive a project even in the short duration of a rotation, which gave me immense confidence not only to structure problems but also to communicate ideas. Thank you so much for your patience, guidance and encouragement through the years. I would like to thank Prof. Hernday for his insights. It has been great working on collaborative projects with you and learning from you. I remember our meeting during the visitation week at UC Merced and your enthusiasm when you first introduced transcriptional regulation and its role in phenotypic switching. This was a key motivation for my work, and I hope to carry the same enthusiasm for science in my future endeavors. I would like to thank Prof. Sindi for her guidance on structuring problems and scientific presentation. Your lecture was one of the first ones I had the opportunity to attend, and I've always been in awe of your ability to make even the complex of problems more approachable. Thank you for your guidance and feedback and helping critically assess the assumptions and limitations. I hope to continue learning from you all and continue collaborating on exciting projects in the future.

I would also like to thank my collaborators, Prof. Richard Bennett and Prof. Nestor Oviedo for the exciting projects. I am very grateful to UC Merced, the QSB department, especially, the department chairs (Prof. Chris Amemiya and Prof. Fred Wolf) and coordinators (Joy Sanchez Bell and Jan Zarate) for their support. I am always grateful to my peers, lab members past and present. Thanks to all my previous lab members who've guided me though stressful times. Thank you Priyanka, Diana, Ashley, Melanie, Akshay, Thad, Rhondene, Craig and Mohammad. I'd also like to thank my peers who were always open to help problem solve. Thank you Austin, Andrea, Fatemeh, Pegah, and Sarina. And thanks to the newest lab members Pri, Nora, Tahirah and Chantz who always helped me destress.

While my PhD journey would not have been possible without my advisors, mentors, committee and peers, my path to the beginning of this journey is inconceivable without the support of my family. My personal thanks to my parents who always supported me and believed in me even when I did not. Thanks to my dad, the first scientist I know, for nurturing my passion for science. Thanks to my mom, the first teacher I know, for pushing me when I needed it and encouraging me when I faltered. Thanks to my husband, who has always been a great sounding board, problem solver and for being the first audience to every one of my presentations. Thanks to my in-laws for always being interested in my research and for the conversations about epigenetics and its role in health and medicine. Thanks also to my grandparents for their heartfelt prayers and warm wishes. And lastly, and at times most importantly, thanks to my sister for always keeping it real.

DEEPIKA GUNASEKARAN

Curriculum Vita

EDUCATION

Jacobs University Bremen, Germany ***09/2007 - 08/2010***
Bachelor of Sciences in Bioinformatics and Computational Biology

Albert-Ludwigs-University, Freiburg, Germany ***06/2011 - 12/2013***
Master of Sciences in Bioinformatics and Systems Biology

University of California, Merced, CA, USA ***08/2018 - Present***
PhD in Quantitative and Systems Biology

RESEARCH POSITIONS

Graduate Research, University of California, Merced, CA, USA ***08/2018 - Present***
Advisors: Prof. Clarissa Nobile and Prof. David Ardell

Internship, Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, CA, USA ***06/2019 - 08/2019***
Supervisor: Dr. Ernst Oberortner

Lab Rotation, University of California, Merced, CA, USA ***08/2018 - 01/2019***
Supervisors: Prof. Mark Sistrom and Prof. Gordon Bennett

Research Assistant, Laboratory for Pathogenesis of Clinical Drug Resistance and Persistence, San Diego State University, CA, USA ***05/2016 - 08/2018***
Supervisor: Prof. Faramarz Valafar

Masters Research, Institute for molecular medicine and cell biology, Albert-Ludwigs-University, Freiburg, Germany ***09/2012 - 11/2013***
Supervisor: Supervisor: Prof. Oliver Schilling

Student Research Assistant, Pharmaceutical bioinformatics division, Albert-Ludwigs-University, Freiburg, Germany ***11/2011 - 01/2013***
Supervisor: Prof. Stefan Günther

Bachelors Research, Max Planck institute for marine microbiology, Bremen, Germany *06/2009 - 06/2010*
Prof. Frank Oliver Glöckner

PUBLISHED WORK

Elghraoui, A., Gunasekaran, D., Ramirez-Busby, S. M., Bishop, E., & Valafar, F. (2022). Hybran: Hybrid Reference Transfer and ab initio Prokaryotic Genome Annotation. bioRxiv, 2022-11.
(doi: https://doi.org/10.1101/2022.11.09.515824)

Modlin, S.J., Elghraoui, A., Gunasekaran, D., Zlotnicki, A.M., Dillon, N.A., Dhillon, N., Kuo, N., Robinhold, C., Chan, C.K., Baughn, A.D. and Valafar, F., 2021. Structure-Aware Mycobacterium tuberculosis Functional Annotation Uncloaks Resistance, Metabolic, and Virulence Genes. Msystems, 6(6), pp.e00673-21.
(doi: https://doi.org/10.1128/mSystems.00673-21)

Heredia, M. Y., Gunasekaran, D., Ikeh, M. A., Nobile, C. J., & Rauceo, J. M. (2020). Transcriptional regulation of the caspofungin-induced cell wall damage response in Candida albicans. Current Genetics, 1-10.
(doi: https://doi.org/10.1007/s00294-020-01105-8)

Heredia, M. Y., Ikeh, M. A., Gunasekaran, D., Conrad, K. A., Filimonava, S., Marotta, D. H., Nobile, C.J. & Rauceo, J. M. (2020). An expanded cell wall damage signaling network is comprised of the transcription factors Rlm1 and Sko1 in Candida albicans. PLoS Genetics, 16(7), e1008908.
(doi: https://doi.org/10.1371/journal.pgen.1008908)

Petrera, A., Gassenhuber, J., Ruf, S., Gunasekaran, D., Esser, J., Shahinian, J.H., Hübschle, T., Rütten, H., Sadowski, T. and Schilling, O., 2016. Cathepsin A inhibition attenuates myocardial infarction-induced heart failure on the functional and proteomic levels. Journal of Translational Medicine, 14(1), p.153.
(doi: https://doi.org/10.1186/s12967-016-0907-8)

Gunasekaran, D.*, Videm, P.*, Schröder, B., Mayer, B., Biniossek, M.L. and Schilling, O., 2014. Automated peptide mapping and protein-topographical annotation of proteomics data. BMC Bioinformatics, 15(1), p.207.
$*$ - Co-first authors

(doi: https://doi.org/10.1186/1471-2105-15-207)

PATENTS

Valafar, F., Modlin, S. J., Radecke, S. M., Westin, E., & Gunasekaran, D. (2023). U.S. Patent Application No. 17/914,662.

PRESENTATIONS

Gunasekaran, D., Perry, A. M., Kaur, J., Ardell, D. H., Nobile, C. J. Evolutionary mechanisms governing biofilm network rewiring in the Candida species. Candida and Candidiasis, 2023. (Poster)

Gunasekaran, D., Elghraoui, A., Modlin, S. J., Jamshidi, N. and Valafar, F. Constructing a genome-scale model of metabolism and gene expression for Mycobacterium tuberculosis. ASM TB: Past, Present and Future, 2017. (Poster)

Gunasekaran, D. Role of enhancers and super-enhancers in determining Candida albicans cell fate. Molecular and Cell Biology Seminar – University of California, Merced. October 14, 2021. Merced, California. (Talk)

Gunasekaran, D. Workflow visualization of synthetic DNA constructs. Synthetic Biology Seminar – Joint Genome Institute. August 8, 2019. Walnut Creek, California. (Talk)

AWARDS

Student Success Award for Poster Presentation at American Society for Microbiology Conference on Tuberculosis in 2017

National Research Traineeship for Intelligent Adaptive Systems by National Science Foundation in 2018

National Research Traineeship for Interdisciplinary Computational Graduate Education Program by National Science Foundation in 2019

School of Natural Sciences Summer Teaching Assistant Top-off Fellowship by UC Merced in 2019

UC President's Lindau Nobel Laureate Fellow in 2021

Abstract

Evolution of the gene regulatory network controlling biofilm formation in *Candida* species

by

Deepika Gunasekaran

Doctor of Philosophy in Quantitative and Systems Biology
University of California, Merced

Professor Clarissa Nobile, Advisor
Professor David Ardell, Co-advisor

Gene expression is one of the most fundamental processes in a cell, allowing genetic information to be processed into functional products. Regulation of gene expression is determined by the underlying gene regulatory networks (GRNs). We can think of a GRN as a directed network with hub nodes denoting master regulators, which are DNA-binding proteins and target nodes denoting downstream genes whose expression is modulated. The directed network edges denote strength of the interaction between the master regulators and downstream target genes. High-throughput genomic and epigenomic data is used to construct GRNs and to estimate the robustness of these networks. In this study, we used the GRN underlying the formation of complex multicellular structures called biofilms, in the yeast species *Candida albicans*, to understand the mechanisms of GRN divergence between species and variation within species. Biofilm formation has evolved multiple times in the fungal tree of life and the ability to form biofilms is highly varied across species and strains. To study the effects of these variations on components of the GRN, we developed three standalone bioinformatic tools. The first tool was developed to infer the structure of GRNs across *Candida* species by identifying and annotating binding loci of master regulators. The second was used to identify mutations in the *C. albicans* population and the effect of these mutations on the network components. The third tool was used to estimate the evolutionary forces acting on the components of the biofilm GRN. Using these tools, we inferred the sources of genetic variations in the biofilm regulatory network components. We found that the motif preferences of the DNA-binding proteins are conserved across large evolutionary distances but their interaction with target genes is highly divergent. This is driven in part by mutations resulting in gains and losses of genomic regions where the DNA-binding proteins preferentially bind. Furthermore, mutations accumulate in segments of DNA-binding proteins that are required to interact with other hub proteins in the network. This affects the overall structure of the network both within and between species.

Chapter 1

Introduction

1.1 Gene regulatory networks

Gene expression is one of the most fundamental processes in a cell, allowing genetic information to be processed into functional products. The regulation of gene expression is determined by underlying gene regulatory networks (GRNs). A GRN is a system of interacting genes and regulatory elements. We can think of a GRN as a directed network (examples shown in Figure 1.1 and Figure 1.2) with hub nodes denoting DNA-binding proteins (also known as transcription factors (TFs)), which determine if and when downstream genes are expressed. Gene expression is determined by where the TFs bind in the genomic DNA (*cis*-regulatory regions in the genome), and what downstream genes they regulate (hereon referred to as target genes). For this study, we define the *cis*-regulatory region as the upstream (5') intergenic region of a protein-coding gene. Gene expression can also be regulated by regions such as the downstream (3') region, introns or extant enhancers (even at a distance greater than 10000 bases from the gene) (Blackwood and Kadonaga; 1998), especially in higher eukaryotes. Other factors such as chromatin structure, accessibility (Klemm et al.; 2019) and regulatory RNAs (Morris and Mattick; 2014) also affect gene expression. In this study, we use a simplified model of the GRN, comprising three components, TFs, target genes and *cis*-regulatory regions, to study the evolution of the biofilm regulatory network in *Candida albicans*.

Evolutionary changes in the components of the GRN are known to alter the phenotype encoded by the GRN (Schember and Halfon; 2022). Phenotypic diversity can be generated by differences in gene content, specifically in protein-coding genes and/or differences in gene expression. For example, studies have shown that gene loss (Helsen et al. (2020)) and loss-of-function variants (Albalat and Cañestro (2016)) increase phenotypic variability, on which selection can act, to increase the fitness of the organism under different environmental conditions. Mutations in coding regions, however, are not sufficient to explain profound phenotypic differences between species

(King and Wilson; 1975) and gene expression divergence (Jacob and Monod; 1961). Nevertheless, we know more about interspecific fixed differences (hereon referred to as differences) and intraspecific genetic variants (hereon referred to as variants), in protein-coding regions compared to *cis*-regulatory regions. Extensive genetic variation in *cis*-regulatory regions has been observed, however, with evidence for both the conservation of transcription factor binding sites (TFBSs) and the rapid divergence of *cis*-regulatory elements (Wray et al. (2003); Hahn (2007); Siepel and Arbiza (2014)). The studies mentioned above focus on identifying evolutionary constraints on protein-coding genes and *cis*-regulatory elements independently, however, knowledge of TF-target gene relationship can be leveraged to better understand divergence of a GRN. Studies that use such prior knowledge uncovered coevolution in modules of regulatory regions and TFs (Gordon and Ruvinsky; 2012), evolutionary path of divergence of a *cis*-regulatory module (Britton et al.; 2020), and mutational compensation (Landry et al.; 2005).

Extending on the principle of coanalysing variants in *cis*-regulatory regions and TFs, if we know the underlying GRN, we can also assess constraints posed on the *cis*-regulatory region as a function of the TFs that bind in this region. This essentially provides a view of the effect of network architecture on the evolutionary constraints placed on the network. Changes in GRNs underlie divergence of functional modules in many species. Evolution of new developmental processes driven by changes in GRNs have been observed, such as, for example, in flagellar organization in bacteria (McAdams et al. (2004)). In fungi, GRN rewiring has been shown to underlie changes in regulatory functions from regulating sterol synthesis to regulating filamentation (Maguire et al.; 2014), and from regulating sporulation to regulating biofilm formation (Nocedal et al.; 2017). Additionally, GRN rewiring in fungi underlies differences in galactose metabolism (Dalal et al.; 2016), and purine catabolism (Tebung et al.; 2016). In invertebrates, GRN rewiring is known to alter gene expression at the developmental stage in sea urchins (Israel et al.; 2016), and is responsible for pigment differences in *Drosophila* (Massey and Wittkopp; 2016). In mammals, GRN rewiring underlies the differentiation of embryonic stem cells in mice (Zhou et al.; 2007) and humans (Kunarso et al.; 2010; Tsankov et al.; 2015). Changes in GRNs can be a consequence of mutations in the TFs that alter their binding specificity or mode of regulation or of mutations in *cis*-regulatory regions of target genes. Consequentially, mutations in *cis*-regulatory regions can result in losses, gains, changes in binding affinities or functional conversions of binding sites. Although these changes can modify GRNs, divergence of GRNs between species is driven primarily by gene duplication (Teichmann and Babu; 2004; Voordeckers et al.; 2015). For example, a gene duplication event and subsequent divergence through neofunctionalization or subfunctionalization of a TF, can result in the generation of paralogs that respond to different environmental signals or in the regulation of different target genes (Teichmann and Babu (2004); Gu et al. (2005); Wohlbach et al. (2009); Ohno (2013); Pougach et al. (2014); Voordeckers et al. (2015); Gu et al. (2004)). TF duplication events can undergo different fates, all of which change the evolutionary landscape of the GRN, including: (i) subfunctionalization, resulting in specialized function of the ancestral gene and its paralog through segregation of target genes between the two copies (Voordeckers et al. (2015)), (ii) redundancy, when both copies of the TF regulate the same genes (De Smet and Van de Peer (2012)), and (iii) neofunctionalization, where the paralog regulates genes not previously regulated by the ancestral TF (Voordeckers et al. (2015)).

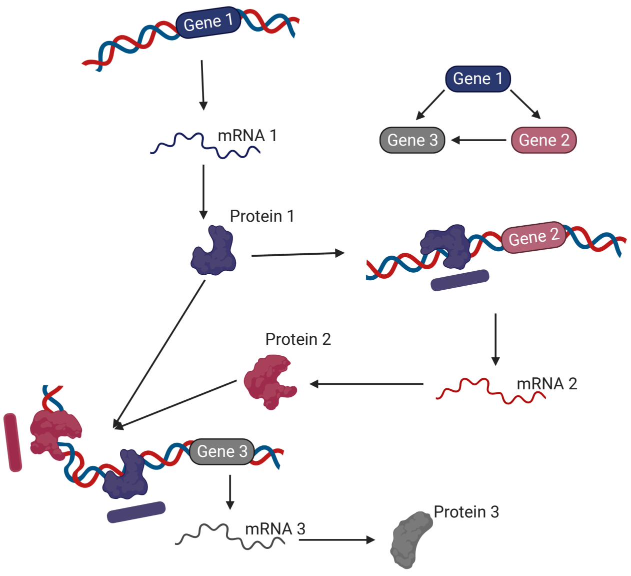


Figure 1.1: **Components of a gene regulatory network**
Gene regulatory networks are comprised of protein-coding genes translated to DNA-binding proteins called TFs that bind to DNA in a sequence-specific manner within *cis*-regulatory regions to regulate transcription of the target genes. Binding sites preferred by each TF is indicated by bars below the DNA, colored by the corresponding TF. In this example, gene 1 regulates gene 2 and protein 1 and protein 2 bind in the intergenic region of gene 3, to regulate this gene in a cooperative manner. Other known elements of GRNs, such as regulatory RNAs are not included in the figure.

1.2 Motivation and study aim

The overall focus of my research is to understand the mechanisms of intraspecific and interspecific evolution of transcriptional networks using the *C. albicans* biofilm network as a model. Although various mechanisms for GRN diversification have been observed and studied in isolation, the effect of these variants in the context of the underlying network structure is still relatively unknown. Elucidating the divergence of network structure requires experimental data from related species, measured for the same phenotype. This data was recently published by Mancera et al. (2021) for the biofilm regulatory networks of several *Candida* species. We reanalysed this data to infer the network structure across species. Additionally, predicting the effect of *cis*-regulatory variants is challenging due to lack of functional annotation of *cis*-regulatory elements. Studies on the evolution of *cis*-regulatory elements have traditionally assumed that any change in the TFBS has functional consequences, however, the effects of natural variants on binding affinity and resulting change in structure of the GRN have not been studied (Hahn (2007); Siepel and Arbiza (2014); Wray et al. (2003)). In this study, we identified TF binding sites across species and used this information to obtain high confidence binding motifs, which were then used to predict the effects of variants. One limitation to note, is that we do not account for other regulatory effects, such as chromatin structure and accessibility.

Studies have also reported interspecific gains and losses of TFBSs, but the frequency of occurrence of these events within a species, mechanisms driving these events and their roles in GRN evolution are unknown (Bradley et al. (2010); Bullaughey (2011); Doniger and Fay (2007); He et al. (2011); Moses et al. (2006); Krieger et al. (2022)). We identify TFBS gain and loss in *C. albicans* for the biofilm regulators. Additionally, previous studies exploring genomic variation of *cis*-regulatory regions have focused on single nucleotide polymorphisms (SNPs) and variations that affect individual TF binding sites. In this study, we expand our understanding of mechanisms driving variations in *cis*-regulatory regions by including larger structural variations (SVs) and studying their relative contribution(s) to the disruption of these regulatory elements. We also developed three tools to identify SNPs and SVs, identify binding sites and measure molecular evolution statistics.

1.3 *Candida albicans* as a model organism

1.3.1 *C. albicans* is a biologically relevant model

Candida species are commensal fungi of the skin, gastrointestinal and genitourinary tracts of humans, but they are also opportunistic pathogens capable of causing superficial and systemic infections under specific host conditions (Silva et al. (2009)). *Candida* species, in particular *C. albicans*, are the most common fungal species isolated from hospital settings (Perlroth et al. (2007)). One significant virulence trait of the *Candida* species is the ability to form multicellular microbial communities called biofilms (Costerton et al. (1995); Silva et al. (2009)). *C. albicans*, interestingly, is also the most robust biofilm former in this genus (Hawser and Douglas (1994); Kuhn et al. (2002); Silva et al. (2009)). The capacity of *C. albicans* to cause disease is due, in part, to its ability to form highly structured biofilms that confer resistance to both physical and chemical stresses (Mayer et al. (2013); Nobile and Johnson (2015);

Kaur and Nobile (2023)). The presence of a *C. albicans* biofilm infection is correlated with a significant decrease in the efficacy of antifungal drug treatment as well as with an increase in the associated morbidity and mortality rates of the patient (Bizerra et al. (2008); Cauda (2009); Rajendran et al. (2016)). Biofilm formation is a highly regulated process in *C. albicans* and the biofilm regulatory network in *C. albicans* is an ideal model system to study the evolution of transcriptional networks since this network is well established, recently evolved, and seemingly less complex relative to known regulatory networks in higher eukaryotes.

1.3.2 Biofilm formation is a well-studied phenotype in *C. albicans*

Biofilm formation has evolved multiple times in the fungal tree of life (Desai et al.; 2014; Naranjo-Ortiz and Gabaldón; 2020). Among the most commonly known biofilm forming fungi are *Aspergillus fumigatis* and *Cryptococcus neoformans* (Fanning and Mitchell; 2012). These robust biofilm-formers are also clinically relevant pathogens. The biofilm regulatory network in *C. albicans* was first described by Nobile et al. (2012) as a network comprised of six "master" regulators (Bcr1, Brg1, Efg1, Ndt80, Rob1 and Tec1) and approximately 1000 downstream target genes (Figure 1.2) (Nobile et al. (2012)). This network was later expanded to include three additional regulators (Flo8, Gal4 and Rfx2), which play distinct roles in different temporal stages of biofilm formation (Fox et al. (2015)). Nobile et al. (2012) also hypothesized that this network is recently evolved, since upregulated genes under biofilm forming conditions emerged relatively recently in the *Ascomycota* phylum, after the diversification of *Schizosaccharomyces pombe* and *Neurospora crassa* (Nobile et al. (2012)). A recent study showed that despite conservation of binding specificities of most of the biofilm regulators across species closely related to *C. albicans*, the targets regulated by these TFs are diverse (Mancera et al. (2021)). While Mancera et al. (2021) focused on interspecies rewiring of the biofilm regulatory network, other studies have demonstrated that biofilm formation is highly varied among *C. albicans* strains (Hawser and Douglas (1994); Huang et al. (2019); Jain et al. (2007); Kuhn et al. (2002); Li et al. (2003); Pujol et al. (2015); Soll and Daniels (2016); Villar-Vidal et al. (2011)). Notably, Huang et al. (2019) observed that the downstream effects of deletion of the biofilm regulators is dependent on their genetic backgrounds (Huang et al. (2019)). In addition to phenotypic characterization of *C. albicans* clinical and environmental strains, studies have also looked at the population structure and genomic diversity driven by SNPs and copy number variants (Bensasson et al. (2019); Hirakawa et al. (2015); Ropars et al. (2018)). Hirakawa et al. (2015) identified that genome-wide heterozygosity is correlated with higher fitness and hypothesized that adaptive genomic changes in clinical isolates are enriched for genes associated with virulence and the host response (Hirakawa et al. (2015)). Ropars et al. (2018) showed that a subspecies of the *C. albicans* clade, *Candida africana*, has undergone extensive pseudogenization in genes associated with virulence and morphogenesis, possibly contributing to its restriction to a specific host environmental niche (Ropars et al. (2018)). Together, these studies suggest that different isolates of *C. albicans* likely have differences in biofilm regulatory network components, affecting their abilities to form biofilms, and consequently leading to changes in their virulence properties in the host. Additionally, these studies also suggest that *cis*-regulatory variants and binding site gains and losses are likely mechanisms of evolution of this regulatory network and play roles in co-opting novel genes into the network. In general, the study of population genomic variants affecting *cis*-regulatory grammar and the identification of conserved modules in the biofilm network will be important information to guide the development of future therapeutic targets against biofilm formation that could be functional across *Candida* species.

1.4 Genetic diversity in yeasts

The monophyletic kingdom of fungi comprises nine phyla, of which *Ascomycota* is the largest and most well studied (Naranjo-Ortiz and Gabaldón (2019); Schoch et al. (2009)). *Ascomycetes* are genotypically and phenotypically diverse, ranging from simple yeasts to highly filamentous fungi (Hane et al. (2011); Naranjo-Ortiz and Gabaldón (2019)). *Ascomycota* consists of three subphyla *Saccharomycotina* (budding yeasts), *Pezizomycotina* (filamentous yeasts) and *Taphrinomycotina* (fission yeasts) (Shen et al. (2020)). The *Saccharomycotina* subphylum includes the model yeast species *Saccharomyces cerevisiae*, and the most relevant human fungal pathogens of the *Candida* genus. Studies have shown that this subphylum has high evolutionary rates, and that diversity is driven primarily by reductive genome evolution (Shen et al. (2018, 2020)). Reductive modes of diversification in budding yeasts have been demonstrated by comparing their metabolic capabilities with filamentous yeasts in the *Ascomycota* phylum. Although the most studied budding yeast, *S. cerevisiae*, displays significantly reduced metabolic capabilities compared to filamentous yeasts, *C. albicans* displays higher metabolic capabilities than average in this subphylum (Shen et al. (2018)). This suggests that robustness is likely maintained in *C. albicans* due to diverse and/or unpredictable environmental conditions (Félix and Wagner (2008); Mendonça et al. (2011)). Additionally, preferential maintenance of robustness and modularity in pathways and networks might also drive maintenance of the size of these networks and genomes (Félix and Wagner (2008); Kitano (2004)).

Studies of *Ascomycota* species have also shown that the rate of gene expression divergence is higher than sequence evolution between species and this has predominantly been attributed to large-scale genomic changes such as whole genome duplication in shared ancestors (Gu et al. (2005); Ihmels et al. (2005)). However, even species with negligible differences in gene content show large differences in gene expression, suggesting that these differences could be a consequence of small-scale *cis*-regulatory fixed differences (Rokas (2022)). Studies have also demonstrated that

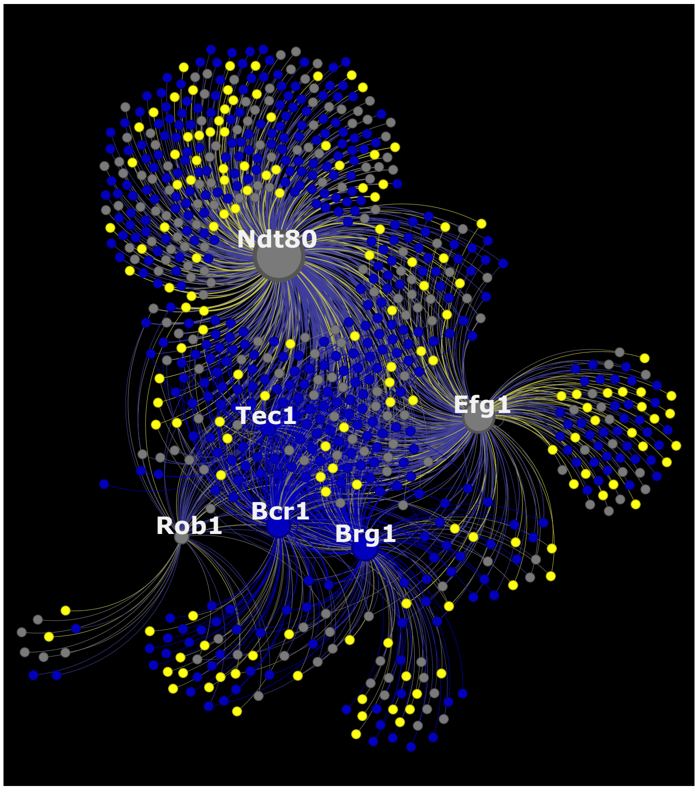


Figure 1.2: **The *C. albicans* biofilm regulatory network**
The biofilm regulatory network identified by Nobile et al. (2012) is comprised of six master transcriptional regulators and approximately 1000 target genes (Nobile et al.; 2012). The regulators are labeled in white and target genes are depicted in blue if they are upregulated, yellow if they are downregulated or gray if no differential expression is observed in biofilm compared to planktonic conditions. The size of the node corresponds to its degree, where the larger nodes are TFs regulating more genes.

genomic evolution in filamentous *Ascomycetes* show conservation of gene content but not gene order or orientation (Hane et al. (2011)), which is likely the consequence of SVs, such as inversions.

Based on previous work, we hypothesize that large scale rewiring of regulatory networks is employed to generate the biofilm regulatory network in *C. albicans*. We observe extensive rewiring in our analysis of synteny across the closely related *Candida* species, *C. albicans*, *C. dubliniensis*, *C. tropicalis* and *C. parapsilosis* (Figure 1.3). Protein-coding genes are highly conserved across these species with more than 80% of genes shared among them. Despite the high conservation of gene content, the gene order is poorly conserved between these species (Figure 1.3), suggesting poor conservation of regulatory elements of these genes.

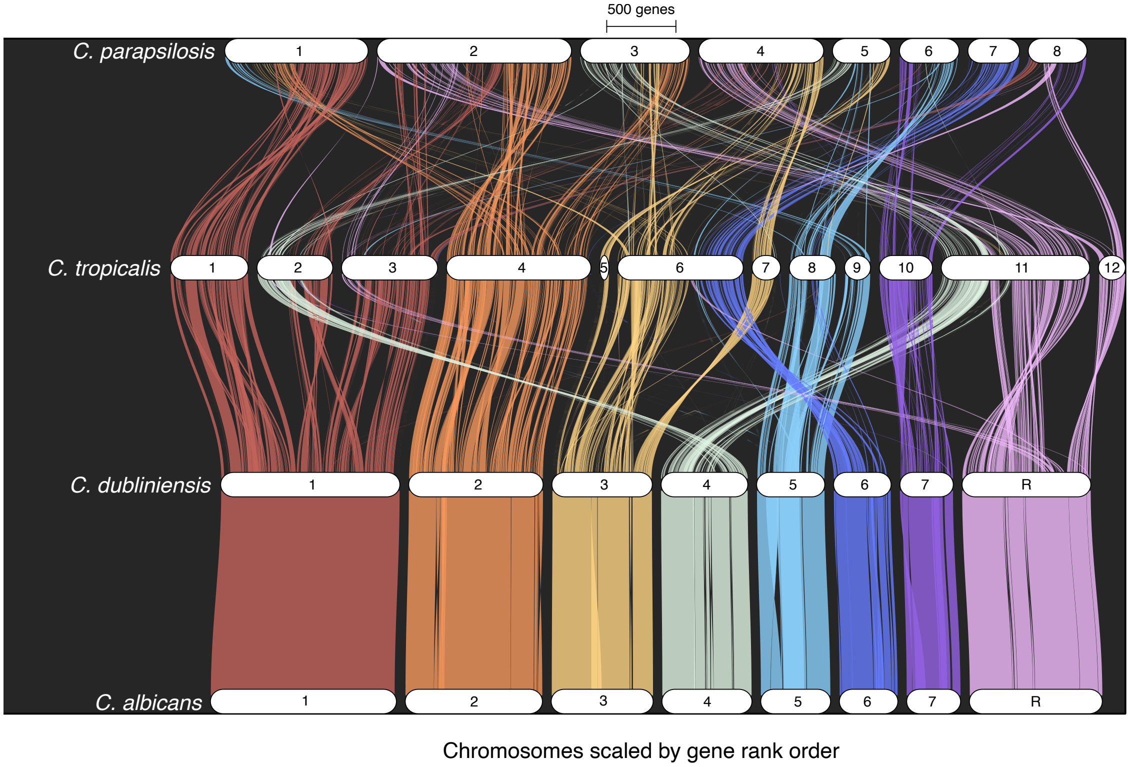


Figure 1.3: **Syntenic relationships between *C. albicans* and its close relatives** Gene order conservation between *C. albicans*, *C. dubliniensis*, *C. tropicalis* and *C. parapsilosis* is depicted in this image. Each line shows a trace of gene order conservation for orthologous genes between these species, and the colors correspond to different chromosomes in *C. albicans*. Orthologous genes were identified using OrthoFinder (Emms and Kelly; 2015, 2019) and gene order conservation was identified using GENESPACE (Lovell et al.; 2022).

1.5 Evolution of gene regulatory networks in yeasts

GRN growth has been attributed primarily to gene duplications (i.e. duplication of TFs and/or target genes in the network) (Teichmann and Babu (2004); Voordeckers et al. (2015)). Simulation of early stages of gene duplication events show maintenance of robustness of these networks (Posadas-García and Espinosa-Soto (2022)), but, over longer evolutionary time, TF-target gene relationships change due to loss of binding sites in promoters of duplicated genes (Ihmels et al. (2005); Maslov et al. (2004);

Conant and Wolfe (2006); Wohlbach et al. (2009); Wapinski et al. (2007)) or gain of motifs of new TFs (Papp et al. (2003); Wohlbach et al. (2009)). This suggests that both subfunctionalization and neofunctionalization contribute to divergence of TF binding (Ihmels et al. (2005); Maslov et al. (2004); Conant and Wolfe (2006); Papp et al. (2003); Evangelisti and Wagner (2004)). These changes result in rapid increased divergence in gene expression in paralogs compared to orthologs (Gu et al. (2002, 2004); Li et al. (2005)).

The studies mentioned above focus on gene expression divergence and GRN rewiring over longer evolutionary time and through the lens of effects of whole genome duplication (WGD) in the *Saccharomyces* clade. However, rewiring has been reported even within species that have diverged from the *Saccharomyces* clade predating the WGD event (Huang et al. (2019); Mancera et al. (2021)). Additionally, recent studies elucidating mechanisms driving expression divergence of target genes of individual TFs, show that *cis*-acting differences are the source of large expression divergence in shorter evolutionary timescales (Siddiq and Wittkopp (2022); Krieger et al. (2022)). These studies also identify *trans*-acting differences and suggest that they are maintained in the population through stabilizing selection (Metzger and Wittkopp (2019); Siddiq and Wittkopp (2022)). Based on this prior work, we hypothesize that *cis*- and *trans*-acting differences contribute to diversity in biofilm formation in the *Candida* species.

The studies mentioned above, focus on individual or on a handful of TFs, and thus provide a focused view on expression divergence driven by *cis*- or *trans*-acting differences in a subset of TFs. In addition to studying these regulatory changes mediated by independent TFs, understanding the changes in binding of TFs that act concertedly in regulating a phenotype, will provide a complementary view on evolutionary constraints acting on GRNs. Theoretical models have shown that phenotypic plasticity is almost always favored for labile traits (Scheiner (1993)) and studies have hypothesized that genes regulated by multiple TFs likely play key roles in the ability of an organism to respond to diverse environments (Promislow (2005); Proulx et al. (2005)). These models suggest that studying the constraints on the number of upstream regulators in a GRN for target genes of interest will shed light on whether plasticity of a trait is favored in a population.

To gain a better understanding of the mechanisms and evolution of GRN divergence between species and variation within species, I specifically address the following questions in my thesis:

i. What is the role of gene duplication in biofilm network evolution across species?

ii. What are the types of genetic variants in the biofilm network components segregating in the *C. albicans* population?

iii. Are *cis*- or *trans*-acting variations more prevalent in the *C. albicans* biofilm network components?

iv. What are the forces of evolution acting on the network components?

v. Is binding affinity, number of binding sites or number of upstream regulators more constrained in the network?

1.6 Bibliography

Albalat, R. and Cañestro, C. (2016). Evolution by gene loss, *Nature Reviews Genetics* 17(7): 379–391.

Bensasson, D., Dicks, J., Ludwig, J. M., Bond, C. J., Elliston, A., Roberts, I. N. and James, S. A. (2019). Diverse lineages of Candida albicans live on old oaks, *Genetics* 211(1): 277–288.

Bizerra, F. C., Nakamura, C. V., De Poersch, C., Estivalet Svidzinski, T. I., Borsato Quesada, R. M., Goldenberg, S., Krieger, M. A. and Yamada-Ogatta, S. F. (2008). Characteristics of biofilm formation by Candida tropicalis and antifungal resistance, *FEMS yeast research* 8(3): 442–450.

Blackwood, E. M. and Kadonaga, J. T. (1998). Going the distance: a current view of enhancer action, *Science* 281(5373): 60–63.

Bradley, R. K., Li, X.-Y., Trapnell, C., Davidson, S., Pachter, L., Chu, H. C., Tonkin, L. A., Biggin, M. D. and Eisen, M. B. (2010). Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species, *PLoS biology* 8(3): e1000343.

Britton, C. S., Sorrells, T. R. and Johnson, A. D. (2020). Protein-coding changes preceded cis-regulatory gains in a newly evolved transcription circuit, *Science* 367(6473): 96–100.

Bullaughey, K. (2011). Changes in selective effects over time facilitate turnover of enhancer sequences, *Genetics* 187(2): 567–582.

Cauda, R. (2009). Candidaemia in patients with an inserted medical device, *Drugs* 69: 33–38.

Conant, G. C. and Wolfe, K. H. (2006). Functional partitioning of yeast co-expression networks after genome duplication, *PLoS biology* 4(4): e109.

Costerton, J. W., Lewandowski, Z., Caldwell, D. E., Korber, D. R. and Lappin-Scott, H. M. (1995). Microbial biofilms, *Annual review of microbiology* 49(1): 711–745.

Dalal, C. K., Zuleta, I. A., Mitchell, K. F., Andes, D. R., El-Samad, H. and Johnson, A. D. (2016). Transcriptional rewiring over evolutionary timescales changes quantitative and qualitative properties of gene expression, *Elife* 5: e18981.

De Smet, R. and Van de Peer, Y. (2012). Redundancy and rewiring of genetic networks following genome-wide duplication events, *Current opinion in plant biology* 15(2): 168–176.

Desai, J. V., Mitchell, A. P. and Andes, D. R. (2014). Fungal biofilms, drug resistance, and recurrent infection, *Cold Spring Harbor perspectives in medicine* 4(10): a019729.

Doniger, S. W. and Fay, J. C. (2007). Frequent gain and loss of functional transcription factor binding sites, *PLoS computational biology* 3(5): e99.

Emms, D. M. and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy, *Genome biology* 16(1): 1–14.

Emms, D. M. and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics, *Genome biology* 20: 1–14.

Evangelisti, A. M. and Wagner, A. (2004). Molecular evolution in the yeast transcriptional regulation network, *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 302(4): 392–411.

Fanning, S. and Mitchell, A. P. (2012). Fungal biofilms, *PLoS pathogens* 8(4): e1002585.

Félix, M.-A. and Wagner, A. (2008). Robustness and evolution: concepts, insights and challenges from a developmental model system, *Heredity* 100(2): 132–140.

Fox, E. P., Bui, C. K., Nett, J. E., Hartooni, N., Mui, M. C., Andes, D. R., Nobile, C. J. and Johnson, A. D. (2015). An expanded regulatory network temporally controls Candida albicans biofilm formation, *Molecular microbiology* 96(6): 1226–1239.

Gordon, K. L. and Ruvinsky, I. (2012). Tempo and mode in evolution of transcriptional regulation, *PLoS genetics* 8(1): e1002432.

Gu, X., Zhang, Z. and Huang, W. (2005). Rapid evolution of expression and regulatory divergences after yeast gene duplication, *Proceedings of the National Academy of Sciences* 102(3): 707–712.

Gu, Z., Nicolae, D., Lu, H. H. and Li, W.-H. (2002). Rapid divergence in expression between duplicate genes inferred from microarray data, *Trends in genetics* 18(12): 609–613.

Gu, Z., Rifkin, S. A., White, K. P. and Li, W.-H. (2004). Duplicate genes increase gene expression diversity within and between species, *Nature genetics* 36(6): 577–579.

Hahn, M. W. (2007). Detecting natural selection on cis-regulatory DNA, *Genetica* 129(1): 7–18.

Hane, J. K., Rouxel, T., Howlett, B. J., Kema, G. H., Goodwin, S. B. and Oliver, R. P. (2011). A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi, *Genome biology* 12: 1–16.

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.

He, B. Z., Holloway, A. K., Maerkl, S. J. and Kreitman, M. (2011). Does positive selection drive transcription factor binding site turnover? a test with Drosophila cis-regulatory modules, *PLoS genetics* 7(4): e1002053.

Helsen, J., Voordeckers, K., Vanderwaeren, L., Santermans, T., Tsontaki, M., Verstrepen, K. J. and Jelier, R. (2020). Gene loss predictably drives evolutionary adaptation, *Molecular biology and evolution* 37(10): 2989–3002.

Hirakawa, M. P., Martinez, D. A., Sakthikumar, S., Anderson, M. Z., Berlin, A., Gujja, S., Zeng, Q., Zisson, E., Wang, J. M., Greenberg, J. M. et al. (2015). Genetic and phenotypic intra-species variation in Candida albicans, *Genome research* 25(3): 413–425.

Huang, M. Y., Woolford, C. A., May, G., McManus, C. J. and Mitchell, A. P. (2019). Circuit diversification in a biofilm regulatory network, *PLoS pathogens* 15(5): e1007787.

Ihmels, J., Bergmann, S., Gerami-Nejad, M., Yanai, I., McClellan, M., Berman, J. and Barkai, N. (2005). Rewiring of the yeast transcriptional network through the evolution of motif usage, *Science* 309(5736): 938–940.

Israel, J. W., Martik, M. L., Byrne, M., Raff, E. C., Raff, R. A., McClay, D. R. and Wray, G. A. (2016). Comparative developmental transcriptomics reveals rewiring of a highly conserved gene regulatory network during a major life history switch in the sea urchin genus Heliocidaris, *PLoS biology* 14(3): e1002391.

Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins, *Journal of molecular biology* 3(3): 318–356.

Jain, N., Kohli, R., Cook, E., Gialanella, P., Chang, T. and Fries, B. (2007). Biofilm formation by and antifungal susceptibility of Candida isolates from urine, *Applied and environmental microbiology* 73(6): 1697–1703.

Kaur, J. and Nobile, C. J. (2023). Antifungal drug-resistance mechanisms in Candida biofilms, *Current Opinion in Microbiology* 71: 102237.

King, M.-C. and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees: Their macromolecules are so alike that regulatory mutations may account for their biological differences, *Science* 188(4184): 107–116.

Kitano, H. (2004). Biological robustness, *Nature Reviews Genetics* 5(11): 826–837.

Klemm, S. L., Shipony, Z. and Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome, *Nature Reviews Genetics* 20(4): 207–220.

Krieger, G., Lupo, O., Wittkopp, P. and Barkai, N. (2022). Evolution of transcription factor binding through sequence variations and turnover of binding sites, *Genome Research* 32(6): 1099–1111.

Kuhn, D., Chandra, J., Mukherjee, P. and Ghannoum, M. (2002). Comparison of biofilms formed by Candida albicans and Candida parapsilosis on bioprosthetic surfaces, *Infection and immunity* 70(2): 878–888.

Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H. and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells, *Nature genetics* 42(7): 631–634.

Landry, C. R., Wittkopp, P. J., Taubes, C. H., Ranz, J. M., Clark, A. G. and Hartl, D. L. (2005). Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of Drosophila, *Genetics* 171(4): 1813–1822.

Li, W.-H., Yang, J. and Gu, X. (2005). Expression divergence between duplicate genes, *TRENDS in Genetics* 21(11): 602–607.

Li, X., Yan, Z. and Xu, J. (2003). Quantitative variation of biofilms among strains in natural populations of Candida albicans, *Microbiology* 149(2): 353–362.

Lovell, J. T., Sreedasyam, A., Schranz, M. E., Wilson, M., Carlson, J. W., Harkess, A., Emms, D., Goodstein, D. M. and Schmutz, J. (2022). GENESPACE tracks regions of interest and gene copy number variation across multiple genomes, *Elife* 11: e78526.

Maguire, S. L., Wang, C., Holland, L. M., Brunel, F., Neuveglise, C., Nicaud, J.-M., Zavrel, M., White, T. C., Wolfe, K. H. and Butler, G. (2014). Zinc finger transcription factors displaced SREBP proteins as the major Sterol regulators during Saccharomycotina evolution, *PLoS genetics* 10(1): e1004076.

Mancera, E., Nocedal, I., Hammel, S., Gulati, M., Mitchell, K. F., Andes, D. R., Nobile, C. J., Butler, G. and Johnson, A. D. (2021). Evolution of the complex transcription network controlling biofilm formation in Candida species, *Elife* 10: e64682.

Maslov, S., Sneppen, K., Eriksen, K. A. and Yan, K.-K. (2004). Upstream plasticity and downstream robustness in evolution of molecular networks, *BMC Evolutionary Biology* 4: 1–12.

Massey, J. and Wittkopp, P. J. (2016). The genetic basis of pigmentation differences within and between Drosophila species, *Current topics in developmental biology* 119: 27–61.

Mayer, F. L., Wilson, D. and Hube, B. (2013). Candida albicans pathogenicity mechanisms, *Virulence* 4(2): 119–128.

McAdams, H. H., Srinivasan, B. and Arkin, A. P. (2004). The evolution of genetic regulatory systems in bacteria, *Nature Reviews Genetics* 5(3): 169–178.

Mendonça, A. G., Alves, R. J. and Pereira-Leal, J. B. (2011). Loss of genetic redundancy in reductive genome evolution, *PLoS computational biology* 7(2): e1001082.

Metzger, B. P. and Wittkopp, P. J. (2019). Compensatory trans-regulatory alleles minimizing variation in TDH3 expression are common within Saccharomyces cerevisiae, *Evolution letters* 3(5): 448–461.

Morris, K. V. and Mattick, J. S. (2014). The rise of regulatory RNA, *Nature Reviews Genetics* 15(6): 423–437.

Moses, A. M., Pollard, D. A., Nix, D. A., Iyer, V. N., Li, X.-Y., Biggin, M. D. and Eisen, M. B. (2006). Large-scale turnover of functional transcription factor binding sites in Drosophila, *PLoS computational biology* 2(10): e130.

Naranjo-Ortiz, M. A. and Gabaldón, T. (2019). Fungal evolution: diversity, taxonomy and phylogeny of the Fungi, *Biological Reviews* 94(6): 2101–2137.

Naranjo-Ortiz, M. A. and Gabaldón, T. (2020). Fungal evolution: cellular, genomic and metabolic complexity, *Biological Reviews* 95(5): 1198–1232.

Nobile, C. J., Fox, E. P., Nett, J. E., Sorrells, T. R., Mitrovich, Q. M., Hernday, A. D., Tuch, B. B., Andes, D. R. and Johnson, A. D. (2012). A recently evolved transcriptional network controls biofilm development in Candida albicans, *Cell* 148(1-2): 126–138.

Nobile, C. J. and Johnson, A. D. (2015). Candida albicans biofilms and human disease, *Annual review of microbiology* 69: 71–92.

Nocedal, I., Mancera, E. and Johnson, A. D. (2017). Gene regulatory network plasticity predates a switch in function of a conserved transcription regulator, *Elife* 6: e23250.

Ohno, S. (2013). *Evolution by gene duplication*, Springer Science & Business Media.

Papp, B., Pál, C. and Hurst, L. D. (2003). Evolution of cis-regulatory elements in duplicated genes of yeast, *TRENDS in Genetics* 19(8): 417–422.

Perlroth, J., Choi, B. and Spellberg, B. (2007). Nosocomial fungal infections: epidemiology, diagnosis, and treatment, *Medical mycology* 45(4): 321–346.

Posadas-García, Y. S. and Espinosa-Soto, C. (2022). Early effects of gene duplication on the robustness and phenotypic variability of gene regulatory networks, *BMC bioinformatics* 23(1): 1–23.

Pougach, K., Voet, A., Kondrashov, F. A., Voordeckers, K., Christiaens, J. F., Baying, B., Benes, V., Sakai, R., Aerts, J., Zhu, B. et al. (2014). Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network, *Nature Communications* 5(1): 4868.

Promislow, D. (2005). A regulatory network analysis of phenotypic plasticity in yeast, *The American naturalist* 165(5): 515–523.

Proulx, S. R., Promislow, D. E. and Phillips, P. C. (2005). Network thinking in ecology and evolution, *Trends in ecology & evolution* 20(6): 345–353.

Pujol, C., Daniels, K. J. and Soll, D. R. (2015). Comparison of switching and biofilm formation between MTL-homozygous strains of Candida albicans and Candida dubliniensis, *Eukaryotic Cell* 14(12): 1186–1202.

Rajendran, R., Sherry, L., Nile, C. J., Sherriff, A., Johnson, E., Hanson, M., Williams, C., Munro, C., Jones, B. and Ramage, G. (2016). Biofilm formation is a risk factor for mortality in patients with Candida albicans bloodstream infection—scotland, 2012–2013, *Clinical Microbiology and Infection* 22(1): 87–93.

Rokas, A. (2022). Evolution of the human pathogenic lifestyle in fungi, *Nature microbiology* 7(5): 607–619.

Ropars, J., Maufrais, C., Diogo, D., Marcet-Houben, M., Perin, A., Sertour, N., Mosca, K., Permal, E., Laval, G., Bouchier, C. et al. (2018). Gene flow contributes to diversification of the major fungal pathogen Candida albicans, *Nature communications* 9(1): 2253.

Scheiner, S. M. (1993). Genetics and evolution of phenotypic plasticity, *Annual review of ecology and systematics* 24(1): 35–68.

Schember, I. and Halfon, M. S. (2022). Common themes and future challenges in understanding gene regulatory network evolution, *Cells* 11(3): 510.

Schoch, C. L., Sung, G.-H., López-Giráldez, F., Townsend, J. P., Miadlikowska, J., Hofstetter, V., Robbertse, B., Matheny, P. B., Kauff, F., Wang, Z. et al. (2009). The Ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits, *Systematic biology* 58(2): 224–239.

Shen, X.-X., Opulente, D. A., Kominek, J., Zhou, X., Steenwyk, J. L., Buh, K. V., Haase, M. A., Wisecaver, J. H., Wang, M., Doering, D. T. et al. (2018). Tempo and mode of genome evolution in the budding yeast subphylum, *Cell* 175(6): 1533–1545.

Shen, X.-X., Steenwyk, J. L., LaBella, A. L., Opulente, D. A., Zhou, X., Kominek, J., Li, Y., Groenewald, M., Hittinger, C. T. and Rokas, A. (2020). Genome-scale phylogeny and contrasting modes of genome evolution in the fungal phylum Ascomycota, *Science advances* 6(45): eabd0079.

Siddiq, M. A. and Wittkopp, P. J. (2022). Mechanisms of regulatory evolution in yeast, *Current Opinion in Genetics & Development* 77: 101998.

Siepel, A. and Arbiza, L. (2014). Cis-regulatory elements and human evolution, *Current opinion in genetics & development* 29: 81–89.

Silva, S., Henriques, M., Martins, A., Oliveira, R., Williams, D. and Azeredo, J. (2009). Biofilms of non-Candida albicans Candida species: quantification, structure and matrix composition, *Sabouraudia* 47(7): 681–689.

Soll, D. R. and Daniels, K. J. (2016). Plasticity of Candida albicans biofilms, *Microbiology and molecular biology reviews* 80(3): 565–595.

Tebung, W. A., Choudhury, B. I., Tebbji, F., Morschhäuser, J. and Whiteway, M. (2016). Rewiring of the Ppr1 zinc cluster transcription factor from purine catabolism to pyrimidine biogenesis in the Saccharomycetaceae, *Current Biology* 26(13): 1677–1687.

Teichmann, S. A. and Babu, M. M. (2004). Gene regulatory network growth by duplication, *Nature genetics* 36(5): 492–496.

Tsankov, A. M., Gu, H., Akopian, V., Ziller, M. J., Donaghey, J., Amit, I., Gnirke, A. and Meissner, A. (2015). Transcription factor binding dynamics during human ES cell differentiation, *Nature* 518(7539): 344–349.

Villar-Vidal, M., Marcos-Arias, C., Eraso, E. and Quindós, G. (2011). Variation in biofilm formation among blood and oral isolates of Candida albicans and Candida dubliniensis, *Enfermedades Infecciosas y Microbiologia Clinica* 29(9): 660–665.

Voordeckers, K., Pougach, K. and Verstrepen, K. J. (2015). How do regulatory networks evolve and expand throughout evolution?, *Current opinion in biotechnology* 34: 180–188.

Wapinski, I., Pfeffer, A., Friedman, N. and Regev, A. (2007). Natural history and evolutionary principles of gene duplication in fungi, *Nature* 449(7158): 54–61.

Wohlbach, D. J., Thompson, D. A., Gasch, A. P. and Regev, A. (2009). From elements to modules: regulatory evolution in Ascomycota fungi, *Current opinion in genetics & development* 19(6): 571–578.

Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V. and Romano, L. A. (2003). The evolution of transcriptional regulation in eukaryotes, *Molecular biology and evolution* 20(9): 1377–1419.

Zhou, Q., Chipperfield, H., Melton, D. A. and Wong, W. H. (2007). A gene regulatory network in mouse embryonic stem cells, *Proceedings of the National Academy of Sciences* 104(42): 16438–16443.

Chapter 2

Transcription factor-target gene conservation across *Candida* species

2.1 Abstract

The ability to form biofilms is highly varied across *Candida* species. This study examines the possible sources of variation in this process. Some *Candida* species translate the CUG codon to serine rather than the more standard leucine. We refer to the *Candida* species with atypical CUG codon translation as “CTG species” and species that use the ancestral genetic code as “non-CTG species”. We found that of the nine known master biofilm regulators in the CTG species *C. albicans*, five are conserved over large evolutionary distances, three are lost or have considerably diverged in non-CTG *Candida* species and one has newly emerged, most likely in the ancestor shared by *C. albicans*, *C. dubliniensis* and *C. tropicalis*. We constructed the biofilm network in *C. dubliniensis*, *C. tropicalis* and *C. parapsilosis* and also expanded the previously identified *C. albicans* biofilm network. We found that binding sites and target genes are only moderately conserved even in closely related CTG species. Several TFs that are conserved across species regulate distinct target genes in each species. In rare occasions, changes in binding preferences drive this change, but more commonly, TFs regulate species-specific functions despite conservation of motif preference. In addition to target genes and binding strengths, network structure is also varied. The *C. albicans* biofilm network is highly intertwined with a large number of target genes regulated by multiple TFs. In *C. albicans*, the coordinated regulation of target genes by multiple TFs is prominent and target genes encoding proteins with annotated functions in host interactions appear to be enriched for such combinatorial interactions.

2.2 Introduction

Candida species are among the most common fungal species that cause bloodstream infections in the United States, which are associated with high mortality rates (Edmond et al. (1999)). *C. albicans* is the predominant pathogen responsible for *Candida* infections followed by *Candida glabrata*, *Candida tropicalis* and *Candida parapsilosis* (Satoh et al. (2009); Gabaldón et al. (2016); Lone and Ahmad (2019)). Other pathogenic *Candida* species include *Candida dubliniensis*, *Candida lusitaniae* and the recently emerged *Candida auris* (Gabaldón et al. (2016); Chow et al. (2020)), which are all causing increased numbers of infections in hospital settings. Biofilm formation has been reported to vary within (Singh et al.; 2019) and between clinical isolates of the aforementioned *Candida* species (Hasan et al.; 2009; Cavalheiro and Teixeira; 2018). Mancera et al. (2021) examined the conservation of the master biofilm regulators in these closely related species and identified that most of the regulators are conserved in the four closely related *Candida* species (*C. albicans*, *C. dubliniensis*, *C. tropicalis* and *C. parapsilosis*), but their roles in biofilm formation have diverged significantly (Mancera et al. (2021)). Although the conservation of some of the master biofilm regulators in closely related *Candida* species is beginning to be explored (Gasch et al. (2004); Nocedal et al. (2017); Mancera et al. (2021); Kaessmann (2010)), the conservation of the targets of these TFs and their *cis*-regulatory elements have not been studied and are likely to play roles in the biofilm-forming abilities of the *Candida* species (Pais et al. (2016); Cavalheiro and Teixeira (2018)). Gene and genome duplication and *de novo* gene formation can also contribute to rewiring of regulatory networks (Teichmann and Babu (2004); Voordeckers et al. (2015)). Expanding the scope of comparative genomic studies to include these elements, would considerably improve our understanding of biofilm formation among *Candida* species.

In order to study the evolution of the biofilm regulatory network across other *Candida* species, the species listed in Table 2.1 were included in this study. *Schizosaccharomyces pombe* was included as an outgroup. These species belong to different clades, indicated in Figure 2.1. This species phylogenetic tree was constructed by reconciling gene trees of 2356 orthogroups identified using OrthoFinder (Emms and Kelly; 2015, 2019). STAG (Emms and Kelly; 2018) and STRIDE (Emms and Kelly; 2017) were used to reconcile the gene trees in OrthoFinder. Branch lengths indicate the average branch lengths for each bipartition in the individual estimate of the species tree from gene trees of orthologs. Most of these *Candida* species belong to the CTG clade, which are known to use an alternative genetic code, where the CUG codon encodes for serine instead of the canonical leucine (Santos et al. (2011)). The CUG codon is unstable in yeast, with multiple, independent reassignments to encode other amino acids instead of the canonical leucine. Krassowski et al. (2018) report the reassignment of this codon in five monophyletic groups (clades) of yeast species and we grouped the *Candida* species in this study based on their CUG codon reassignment and corresponding clade designation (Krassowski et al.; 2018). All species in this study, except for the outgoup *S. pombe*, share a common ancestor with an altered genetic code. *S. cerevisiae*, *C. glabrata*, *C. kefyr* and *C. pelliculosa* belong to the CUG-Leu1 clade, where the tRNA decoding the CUG codon as leucine is modified in some species of this clade through intron expansion and subsequent changes to its secondary structure or complete loss of this tRNA. *C. krusei* belongs to the CUG-Leu2 clade, which is distinguished from the CUG-Leu1 clade, since they do not share a common ancestor. All other *Candida* species in this study belong to the CUG-Ser1 clade, where the CUG codon encodes for serine. Estimates indicate that the CTG clade diverged from the common ancestor of all *Candida* species approximately 150-170 million years ago (MYA) (Pesole et al.; 1995; Massey et al.; 2003). In Figure 2.1, the phylogenetic relationship between the clades indicates that CUG-Leu1 and CUG-Ser1 clades share a common ancestor, which is contradictory to the findings of Krassowski et al. (2018), where CUG-Leu2 and CUG-Ser1 share a common ancestor. This discrepancy is likely due to the inclusion of a fragmented assembly of *C. pelliculosa*. Therefore, when we retained only the species with complete or near-complete (Figure 2.2), contiguous genome assemblies we inferred phylogenetic relationships between the CUG-Leu1, CUG-Leu2 and CUG-Ser1 clades that were consistent with previous findings (Krassowski et al.; 2018).

Opportunistic fungal pathogens are predominant in the CTG clade, likely due to gene family expansion in virulence related genes (Turner and Butler; 2014). In the ancestor shared between *S. cerevisiae* and *C. glabrata*, a whole genome duplication (WGD) event occurred approximately 100 MYA (Wolfe and Shields; 1997; Kellis et al.; 2004). This WGD event resulted in an expansion of transcription factors in *S. cerevisiae* (Wolfe and Shields; 1997), whereas the *C. glabrata* genome is characterised by gene loss and loss of synteny compared to *S. cerevisiae* (Gabaldón et al.; 2013). We follow the commonly used designation for these species consistent with the literature (Wolfe and Shields; 1997; Gabaldón et al.; 2013; Krassowski et al.; 2018). For instance, when referring to only *S. cerevisiae* and *C. glabrata*, we call them “WGD species”, and when referring to all *Candida* species that use the alternative genetic code we call them “CTG species”. Any species not belonging to this clade (including WGD species), we call “non-CTG species”. We also introduce the term “haploid CTG species” to specifically indicate CTG species that are found predominantly as haploids in nature, such as *C. guilliermondii*, *C. lusitaniae*, *C. auris*, *C. haemulonii*, and *C. duobushaemulonii*, and “diploid CTG species” to specifically indicate CTG species that are found predominantly as diploids in nature (Muñoz et al.; 2018).

2.2.1 Ability of *Candida* species to form biofilms

The *Candida* species included in our study vary in cell size, morphology and biofilm structure. We examined the source of divergence of the underlying biofilm network. The phylogenetic relationships of the *Candida* species in our study are depicted

Species	Strain	Accession	Source	Assembly Level	Contigs	Genome Size	Genes
C. albicans	SC5314	GCF_000182965.3	RefSeq	Chromosome	8	14.3 Mb	6030
C. auris	B11245	GCA_008275145.1	GenBank	Complete	7	12.4 Mb	5327
C. dubliniensis	CD36	GCF_000026945.1	RefSeq	Complete	8	14.6 Mb	5843
C. duobushaemulonii	B09383	GCF_002926085.2	RefSeq	Contig	7	12.6 Mb	5173
C. glabrata	BG2	GCA_014217725.1	GenBank	Complete	13	12.7 Mb	5212
C. guilliermondii	ATCC 6260	GCF_000149425.1	RefSeq	Scaffold	9	10.6 Mb	5920
C. haemulonii	CA3LBN	GCA_019332025.1	GenBank	Complete	7	12.6 Mb	5249
C. kefyr	DMKU3-1042	GCF_001417885.1	RefSeq	Complete	8	10.9 Mb	4805
C. krusei	CBS573	GCF_003054445.1	RefSeq	Complete	5	10.8 Mb	5139
C. lusitaniae	P1	GCA_009498055.1	GenBank	Complete	8	12.1 Mb	5936
C. pelliculosa	NRRL Y-366-8	GCA_001661255.1	GenBank	Scaffold	46	14.1 Mb	6421
C. parapsilosis	CDC317	GCF_000182765.1	RefSeq	Contig	8	13 Mb	5853
C. tropicalis	MYA-3404	GCA_017315405.1	GenBank	Scaffold	16	14.5 Mb	6254
S. cerevisiae	S288C	GCF_000146045.2	RefSeq	Complete	16	12.1 Mb	6017
S. pombe	972h	GCF_000002945.1	RefSeq	Chromosome	3	12.6 Mb	5162

Table 2.1: List of species and assemblies used for this study.

in Figure 2.1, which includes an artist rendering (by Laurence Gao) of the biofilm structures of these species. *C. dubliniensis* and *C. tropicalis* biofilms are more similar to *C. albicans* biofilms, consisting of yeast-form, hyphal and pseudohyphal cells, as well as a robust extracellular matrix (Ramage et al. (2001); Zuza-Alves et al. (2017)). *C. parapsilosis* biofilms consist of only yeast-form and pseudohyphal cells, with a diminished extracellular matrix (Lattif et al. (2010)). *C. duobushaemulonii*, *C. haemulonii* and *C. auris* belong to the same clade and biofilms of species in this clade are comprised primarily of a dense network of yeast-form cells, with occasional pseudohyphal cells reported for some strains (Cendejas-Bueno et al. (2012); Ramos et al. (2020); Sherry et al. (2017)). *C. lusitaniae* forms very thin biofilms comprised predominantly of yeast-form cells (Mancera et al. (2021)). *C. guilliermondii* biofilms also consist primarily of yeast-form cells, but under some conditions, this species has been reported to form pseudohyphal cells (Lastauskienė et al. (2015); Marcos-Zambrano et al. (2017)). The non-CTG *Candida* species also form weak biofilms composed of yeast-form cells and in some cases pseudohyphal cells, with the exception of *C. krusei*, which has been reported to form true hyphal cells (Gómez-Gaviria and Mora-Montes (2020)). The ability to form hyphae (filamentation) is a key process in biofilm formation (Nobile and Johnson; 2015). Diploid CTG species are known to filament (Priest and Lorenz; 2015) and form true hyphae, which likely contributes to robust biofilm formation. Although some haploid CTG species have been reported to filament on occasion (Yue et al.; 2018), they are primarily found in the yeast-form (Muñoz et al.; 2018). Additionally, key genes required for filamentation are lost in haploid CTG species (Muñoz et al.; 2018). In WGD species, TF family expansion has been observed (Wolfe and Shields; 1997), followed by rapid divergence and loss of duplicates in some species (Byrne and Wolfe; 2007). Given the genomic divergence (in WGD species) and divergence in biofilm-forming abilities, we would predict that significant rewiring of the underlying biofilm regulatory network is likely.

2.2.2 Genomic variation in *Candida* species

Previous work, comparing simulations to observed WGD events, estimated retention of 8% of duplicated genes and extensive genomic rearrangements driven by reciprocal translocations (Seoighe and Wolfe (1998)). Reciprocal translocations involve exchange of genomic content between non-homologous chromosomes, which might result in disruption of gene regulation in these segments or fusion of regulatory elements to bring these genes under control of other upstream regulators. Studies have also shown that increased rates of genome rearrangements are observed in CTG species and WGD species, compared to non-WGD species (Rajeh et al. (2018)). Additionally, CTG species have been reported to exhibit increased evolutionary rates compared to non-CTG species, suggesting increased genetic diversity in CTG species (Shen et al. (2020)). Even between the closely related CTG species *C. albicans* and *C. dubliniensis*, differences in genomic content have been reported (Jackson et al. (2009)). While expansion in genes encoding virulence factors has been observed in *C.*

albicans, *C. dubliniensis* has reportedly undergone pseudogenization (Jackson et al. (2009)). In summary, variation in gene content, genomic rearrangements and increased evolutionary rates drive genetic diversity, possibly enabling divergence of the biofilm regulatory network.

2.2.3 Gene regulatory network (GRN) evolution in yeasts

Functional divergence of transcriptional regulation, either by loss of *cis*-regulatory elements (Ihmels et al.; 2005), or by handoff of target genes from one regulator to another (Tuch et al.; 2008; Dalal et al.; 2016; Johnson; 2017; Hsu et al.; 2021) has been reported in fungal species. Ihmels et al. (2005) show that loss of a *cis*-regulatory element resulted in emergence of aerobic growth in WGD yeast species (Ihmels et al.; 2005). In this study, loss and gain of a *cis*-regulatory motif was compared between WGD and non-WGD species. The regulators binding to this motif were unknown, therefore inferences cannot be made on whether this *cis*-regulatory change or *trans*-acting variants are the primary driving force of this rewiring. Studies aimed at understanding handoff of target genes between regulators showed that non-functional binding events are an important intermediary step in facilitating the gain of novel regulatory roles of a TF (Hsu et al.; 2021). Furthermore, cooperative binding of regulators in the *cis*-regulatory region also facilitates the handoff of target genes (Tuch et al.; 2008; Baker et al.; 2012; Johnson; 2017). But in order to study cooperative regulation of TFs, we would require knowledge of all upstream regulators and target genes that drive a specific phenotypic response. Since, all the master regulators of biofilm formation have been identified in *C. albicans* and their role in regulating biofilm formation in other *Candida* species have also been studied, the biofilm GRN is a informative, well-described model to study GRN divergence and evolution. Additionally, in a recent study, Mancera et al. (2021) published the data for *in vivo* binding of these master regulators in three other *Candida* species. This helps us compare the divergence of targets and network structure across species directly.

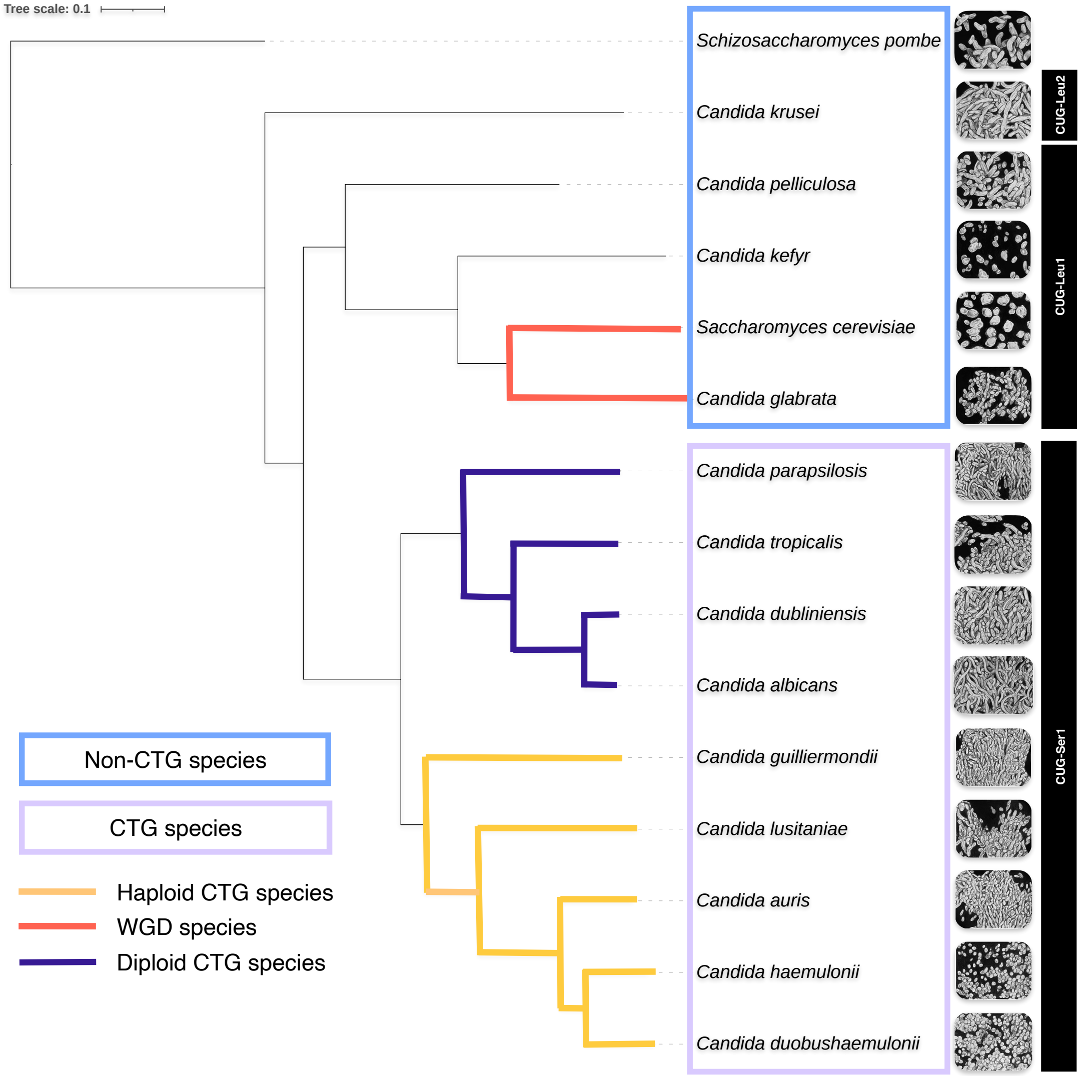


Figure 2.1: **Morphological diversity in *Candida* species.**
Candida species are highly diverse in their morphology and their ability to form biofilms. An artist rendition of the different morphological cell types involved in the biofilms of different *Candida* species as well as the outgroup species *S. cerevisiae* and *S. pombe*. The illustration of biofilms was made by Laurence Gao, an artist in residence in the Nobile lab. The data used to estimate biofilm density in these species was generated by Melanie Ikeh, a former postdoctoral scholar in the Nobile lab. Cell densities in the biofilm communities of these species relative to *C. albicans* is also depicted. The species phylogenetic tree was constructed by reconciling gene trees of 2356 orthogroups identified using OrthoFinder (Emms and Kelly; 2015, 2019). STAG (Emms and Kelly; 2018) and STRIDE (Emms and Kelly; 2017) were used to reconcile the gene trees in OrthoFinder. The branch lengths indicate average branch lengths for each bipartition in the individual estimate of the species tree from gene trees of orthologs.

2.3 Results

2.3.1 Gene loss and duplication of master regulators drive differences in biofilm formation between CTG and non-CTG *Candida* species

Genome assembly and annotation for the 15 species listed in Table 2.1 were obtained as described in subsection 2.5.1. *C. pelliculosa* was removed from further analysis due to its fragmented genome assembly. Orthologous gene relationships were inferred for 14 species, comprising a total of 73499 genes, as described in subsection 2.5.2. Most of these genes (94.3%) were assigned to 6661 orthogroups, among which 2350 orthogroups contained all species and 1579 of these were single-copy orthogroups, with exactly one gene from each species. The presence of orthologs of the nine master regulators across species is depicted in Figure 2.2. These TFs regulate different stages of biofilm formation. Bcr1 regulates adhesion by regulating adhesins *in vitro* and *in vivo* (Nobile et al.; 2006). Brg1 regulates hyphal formation (Cleary et al.; 2012) in response to multiple cues including nutrient limitation and pH (Luo et al.; 2021). *EFG1* is a well studied gene in *C. albicans* (Glazier; 2022) and regulates morphological transition (Do et al.; 2022), hyphal development (Leng et al.; 2001) and extracellular matrix composition (Panariello et al.; 2017). Flo8 interacts with Efg1 *in vivo* to regulate hyphal development and biofilm formation (Fox et al.; 2015). Gal4 and Rfx2 negatively regulate bioiflm formation in intermediary stages (Fox et al.; 2015). Gal4 performs different functions in *C. albicans* compared to *S. cerevisiae* due to a truncated C-terminal activation domain in the CTG species (Choudhury and Whiteway; 2018). Ndt80 regulates meiosis and sporulation in *S. cerevisiae* but has undergone extensive rewiring in *C. albicans* and regulates biofilm formation (Nocedal et al.; 2017). Rob1 regulates filamentation (Glazier et al.; 2017) and the *ROB1* gene in the *C. albicans* reference strain SC5314 is known to contain a dominant allele that modulates biofilm formation and virulence in *in vivo* infection models (Glazier et al.; 2023). Tec1 regulates filamentation and extracellular matrix composition (Panariello et al.; 2017). There is a balance between biofilm regulation and commensalism *in vivo* and Efg1, Brg1 and Rob1 also regulate commensalism (Witchley et al.; 2019). Most biofilm regulators (*BCR1*, *BRG1*, *EFG1*, *FLO8*, *NDT80* and *TEC1*) are conserved across large evolutionary distances. *ROB1* is present only in *C. albicans*, *C. dubliniensis*, and *C. tropicalis*. *GAL4* and *RFX2* are absent in the non-CTG species. Gene duplication events distinct from the WGD duplication were identified in this study and paralogs of *BCR1*, *BRG1*, *FLO8*, *NDT80*, and *TEC1* were present predominantly in the non-CTG species. The paralogs of *BCR1* and *NDT80* can be explained by a single, duplication event that is distinct from the WGD event. *BCR1* duplication was inferred in the ancestor of CUG-Leu1 clade species, *NDT80* duplication was inferred in the ancestor of CUG-Ser1 clade species. A single duplication event was also observed for *TEC1* in *C. glabrata*. The absence of *TEC1* paralog in *S. cerevisiae* could be due to loss of this paralog post-WGD in *S. cerevisiae* or duplication of *TEC1* and subsequent retention of the paralog in *C. glabrata*. On the other hand, multiple duplication events were inferred for *BRG1* and *FLO8* (Figure 2.3). These results suggest that gene loss and duplication of multiple TFs in the non-CTG species could be driving factors leading to differences in the biofilm-forming abilities between the CTG and non-CTG *Candida* species.

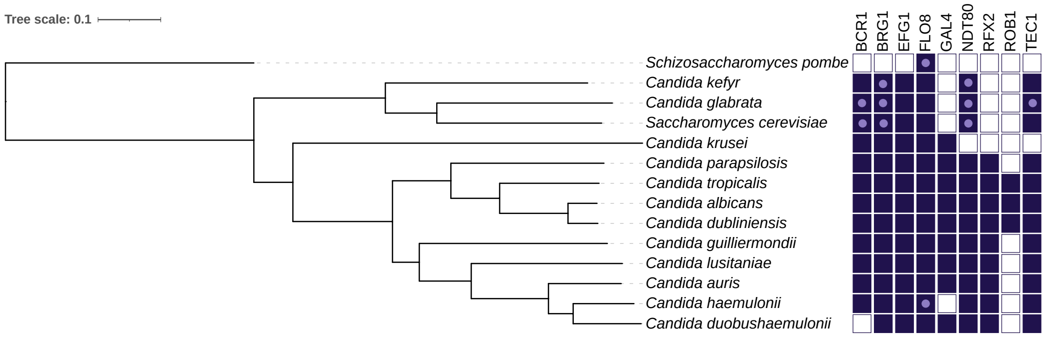


Figure 2.2: **Loss of biofilm master regulators in the non-CTG *Candida* species, *S. cerevisiae* and *S. pombe*.**
Orthologs of the nine *C. albicans* master biofilm regulators were identified using OrthoFinder (Emms and Kelly; 2015, 2019). If an ortholog of the TF is present in a species, it is indicated with a solid blue square, and if a paralog of the TF is also identified in a species, it is indicated with a purple dot within a square. Unfilled squares indicate the absence of an ortholog or a paralog. The species tree was constructed using STAG (Emms and Kelly; 2018) and STRIDE (Emms and Kelly; 2017) in OrthoFinder. The branch lengths indicate average branch lengths for each bipartition in the estimated species tree from gene trees of orthologs.

2.3.2 Genes expressed in mature biofilms are less conserved across large evolutionary distances

We then proceeded to examine if target genes are conserved across the 14 species. To do this, we processed publicly available ChIP-Seq data, measuring *in vivo* binding events of six biofilm regulators in *C. albicans*. We identified 820 putative target genes for Bcr1, 888 for Brg1, 1107 for Efg1, 2073 for Ndt80, 315 for Rob1 and 1099 for Tec1 in *C. albicans*. The orthologous relationships of these target genes were identified across the other 13 species. We then obtained gene expression information from Nobile et al. (2012) and categorized the target genes based on their expression under biofilm relative to planktonic growth conditions in *C. albicans* (Nobile et al. (2012)). Differential gene expression in biofilm growth condition relative to planktonic growth condition was determined by reconciling information from microarray and RNA-Seq assays in Nobile et al. (2012). The proportion of conservation of target

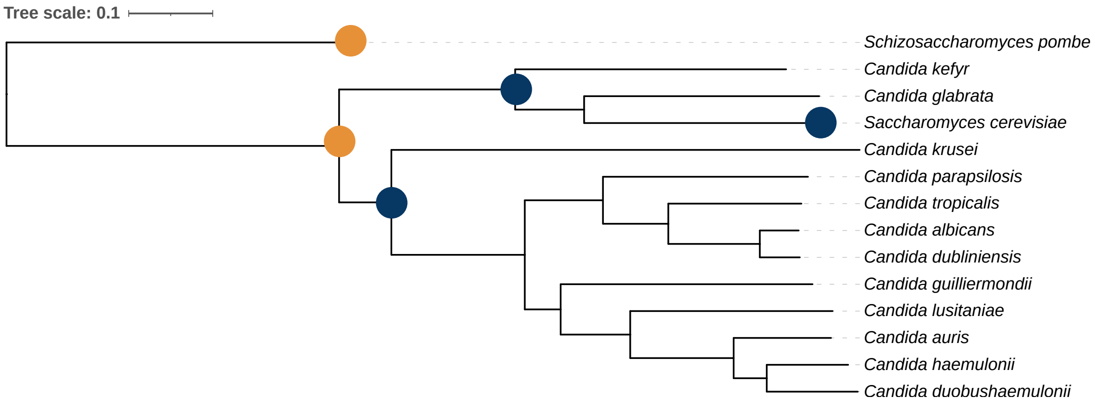


Figure 2.3: **Gene duplication events in *BRG1* and *FLO8*.**
Gene duplication events were inferred using OrthoFinder (Emms and Kelly; 2015, 2019). Blue circles in the phylogenetic tree indicate the nodes at which a *BRG1* duplication event occurred and yellow circles indicate nodes at which *FLO8* duplications occurred. Multiple *BRG1* duplications are identified in the non-CTG *Candida* species. Duplication events are identified in *FLO8* in *S. pombe* and in the ancestor of the CTG and non-CTG species. The branch lengths indicate average branch lengths for each bipartition in the estimated species tree from gene trees of orthologs.

genes in the biofilm network is shown in Figure 2.4. This figure shows that *C. albicans* genes expressed in planktonic conditions (labeled as "Downregulated") are the most conserved across large evolutionary distances and genes expressed in mature biofilms (labeled as "Upregulated") are less conserved, especially in the non-CTG *Candida* species. Genes expressed in planktonic conditions are also significantly more conserved compared to the genes that are not differentially expressed (labeled as "Non-DEG") between the two conditions. Conversely, genes expressed in biofilm condition are significantly less conserved than genes that are not differentially expressed between the two conditions.

In addition to the conservation of target genes categorized by their gene expression, we also verified if target genes of some TFs are more conserved than others. We separated the CTG and non-CTG species for this comparison because target genes, especially those expressed in biofilm conditions are less conserved in the non-CTG species. The comparison of target gene conservation for the six TFs is shown in Figure 2.5. As expected, target genes are less conserved in the non-CTG species compared to the CTG species. In the CTG species, 71%-78% of target genes are conserved, and in the non-CTG species 44%-63% of the target genes are conserved. Rob1 targets are the least conserved in the non-CTG species compared to target genes of other TFs in the biofilm network. We also verified if the proportion of conserved

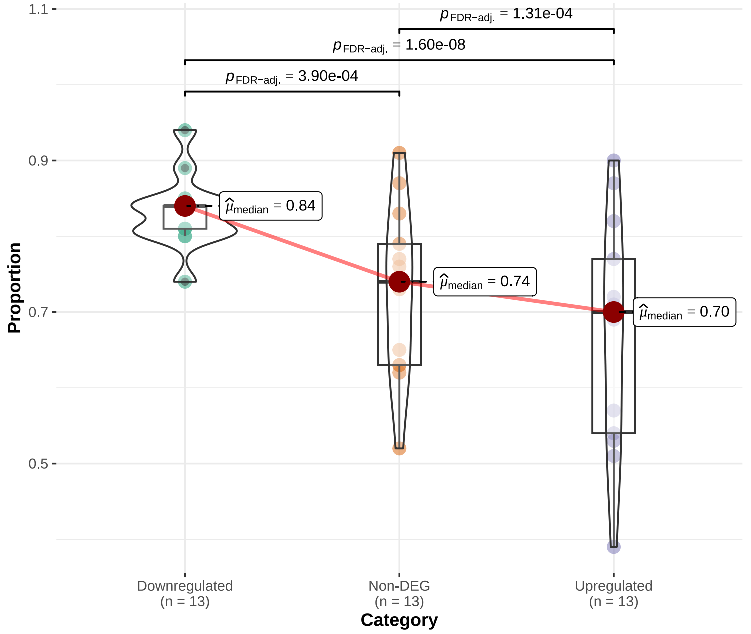


Figure 2.4: **Genes expressed in biofilm condition are significantly less conserved in *Candida* species, *S. cerevisiae* and *S. pombe*.**
Conservation of target genes of the biofilm network based on their expression in planktonic and biofilm conditions (Nobile et al.; 2012). Genes categorized as "Downregulated" are those whose expression is at least 1.5 fold lower in biofilm conditions compared to planktonic conditions. Genes categorized as "Upregulated" are those that are at least 1.5 fold higher in biofilm conditions compared to planktonic conditions and "Non-DEG" are target genes that are not differentially expressed between the two conditions. The statistical significance of differences in gene conservation between these three categories was measured using the Friedman test (Friedman; 1937), a non-parametric test for dependent samples. The $\chi^2_{Friedman}$-value is 19.88. The p-value was adjusted for multiple comparisons using the False Discovery Rate (FDR) and is shown in the plot (Benjamini and Hochberg; 1995).

genes was dependent on the number of TFs regulating the gene and found that there was no significant difference between target genes regulated by only one TF and target genes regulated by all six TFs.

2.3.3 Regulatory role of Ndt80 is the most conserved between *C. albicans* and its close relatives compared to other master regulators

Given the high conservation of target genes in the CTG clade, we wanted to evaluate if binding of the TFs at these genes are also conserved among closely related species.

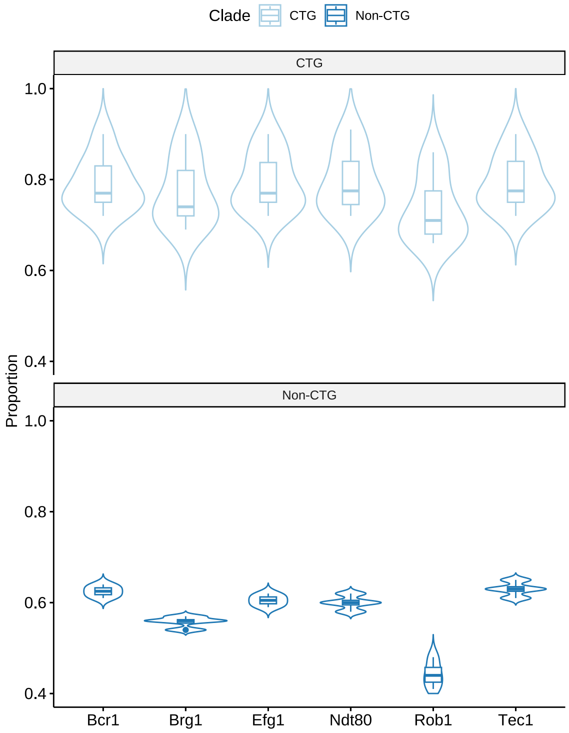


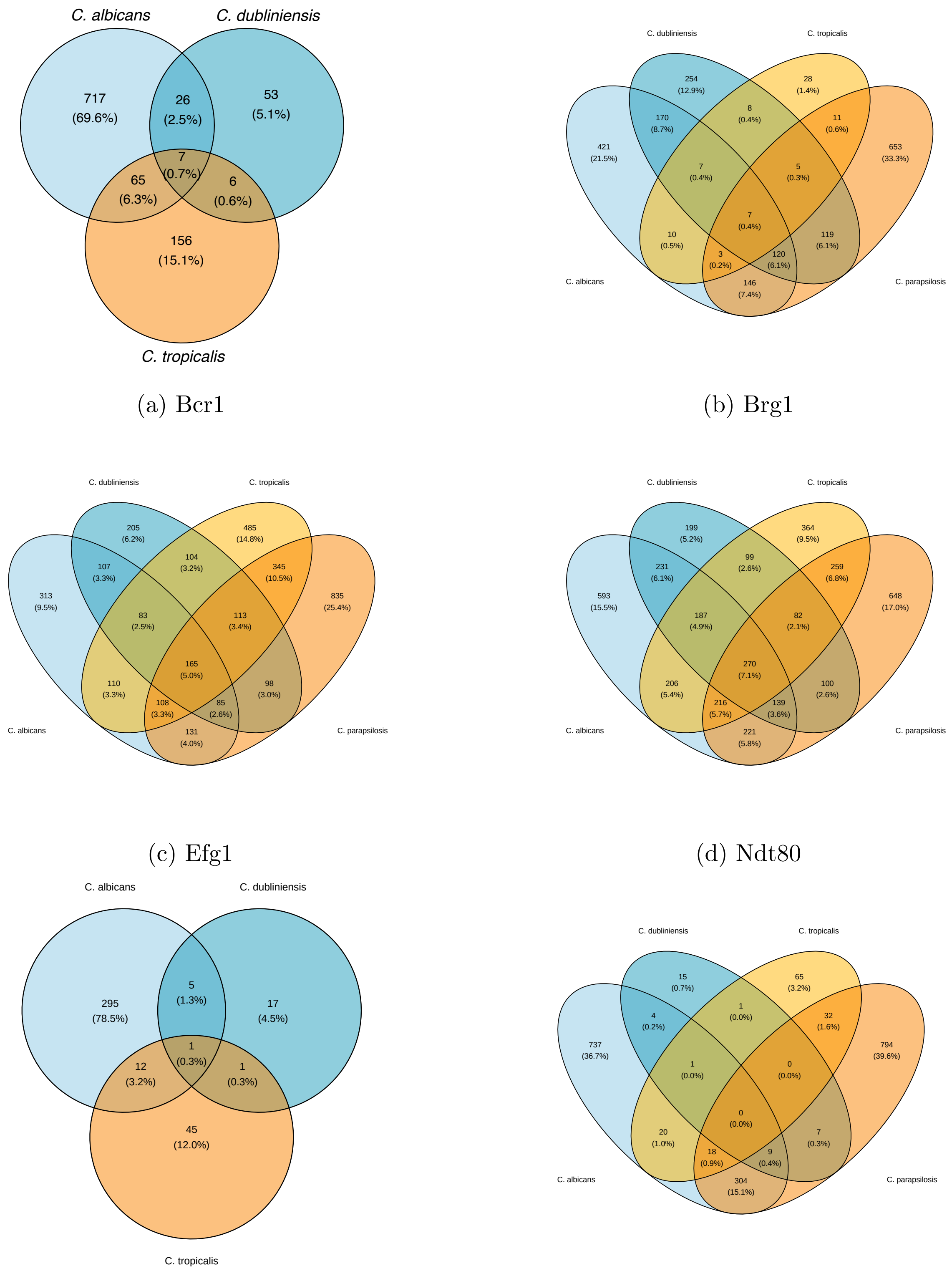
Figure 2.5: **Target genes regulated by Rob1 are less conserved in the non-CTG species.**
Comparison of target gene conservation for each TF between the CTG and non-CTG species. *S. pombe* was removed from this analysis since it is the outgroup. Conservation of targets are higher in the CTG species for all TFs compared to the non-CTG species. Additionally, Rob1 targets are less conserved in the non-CTG species compared to targets of other TFs.

To this end, we compared binding events of the biofilm regulators in *C. albicans*, *C. dubliniensis*, *C. tropicalis* and *C. parapsilosis*. The number of target genes shared across species is shown in Figure 2.6. Ndt80 and Efg1 share the most number of target genes across species, with 270 and 165 target genes, respectively. Bcr1 ChIP-seq data was available only for *C. albicans*, *C. dubliniensis* and *C. tropicalis*. Approximately 90% of Bcr1 targets are species-specific, indicating a possible rewiring of Bcr1 in these three species. Similarly, 69% of Brg1 targets are species-specific. *C. albicans* and *C. dubliniensis* share 170 target genes, similarly 146 target genes are shared by *C. albicans* and *C. parapsilosis*, suggesting that both gains and losses of targets are likely for Brg1. Although a large number of genes were identified as Tec1 targets in *C. albicans* and *C. parapsilosis*, no target genes were common across all four species for this TF. Note that Brg1 and Tec1 are not required for biofilm formation in *C. parapsilosis* (Mancera et al. (2021)). More than 50% of Efg1 targets are species-specific. 345 gene targets are common between *C. parapsilosis* and *C. tropicalis*, and the regulation of these genes by Efg1 was lost in *C. albicans* and *C. dubliniensis*. Conversely, only 107 targets are common between *C. albicans* and *C. dubliniensis*, indicating that a substantial number of ancestral binding events of Efg1 were lost in *C. albicans*. More than 50% of Ndt80 targets are shared between at least two species. Ndt80 also shares the most number of targets between the four species.

We also were interested to know if binding strength is correlated for these TFs between species. In the genes commonly regulated across species, we computed the correlation of cumulative binding strength in the upstream intergenic regions of these genes. The Bayesian Pearson correlation is shown in Figure 2.7. Bayesian Pearson (Ly et al.; 2016) gives the relative likelihood of the observed data under the alternate hypothesis compared to the null hypothesis. In this comparison, the null hypothesis states that the cumulative binding strengths for each TF in the intergenic region of the target genes are not correlated between species. The summarized Bayesian Pearson correlation indicates that Efg1 and Ndt80 binding strength is correlated across species. Cumulative binding strength for Brg1 and Rob1 are most correlated between *C. albicans* and *C. dubliniensis* and binding strength is divergent in the other two *Candida* species for Brg1. There were no common genes regulated by Rob1 between *C. dubliniensis* and *C. tropicalis*.

2.3.4 Binding preferences of biofilm regulators are conserved across diploid CTG species

Given the variation of TF binding and binding strength even in closely related species, we asked whether this variation could be due to changes in binding preference of the TFs. We identified binding motifs for the TFs in *C. albicans*, *C. dubliniensis*, *C. tropicalis* and *C. parapsilosis*. Representative motifs were selected for each TF, prioritizing centrality of the motif in relation to the peak summit identified by MACS2, representation of the motif in binding peaks and information content (IC). Similarity between motifs was estimated using average log-likelihood ratio (ALLR) to combine similar motifs for a TF within species and to compare motifs for the TF between species. The primary motifs for each TF, along with the total IC and percentage of peaks containing the motif is given in Figure 2.8. Bcr1 motifs identified in *C. albicans*, *C. dubliniensis* and *C. tropicalis* share similarity to each other. The ALLR values were computed between *C. albicans* motifs and the other species as described in Section 2.5.6. The ALLR of Bcr1 primary motifs between *C. albicans* and *C. dubliniensis* is 0.996, indicating that the motifs are very similar. The Bcr1 motif in *C. tropicalis* indicates a different nucleotide preference in position four compared to the *C. albicans* motif. Bcr1's binding preference is less conserved between *C. albicans* and *C. tropicalis* (ALLR = 0.486). The primary Bcr1 motif was also identified in sequences retrieved from Brg1. This motif is present in 33.8% of Brg1 sequences in *C.*



(a) Bcr1

(b) Brg1

(c) Efg1

(d) Ndt80

(e) Rob1

(f) Tec1

Figure 2.6: **Conservation of TF-target relationship in four diploid CTG *Candida* species.** Comparison of target genes across species that each TF binds upstream of and regulates. ChIP-Seq data from Mancera et al. (2021) was reanalysed using a custom workflow to obtain binding loci of TFs. This custom ChIP-Seq workflow is described in detail in chapter 4 and was developed to improve sensitivity in identifying binding sites for TFs, especially for single-end and short-reads (approximately, 50 bps). *ROB1* is absent in *C. parapsilosis* and Mancera et al. (2021) were unable to tag Bcr1 in *C. parapsilosis*. Efg1 and Ndt80 share the most target genes across species. Tec1 regulates a large number of species-specific targets. Rob1 and Bcr1 regulate more target genes in *C. albicans* than in the other species.

albicans, 23.2% of sequences in *C. dubliniensis*, 57.6% of sequences in *C. tropicalis*, and 10.4% of sequences in *C. parapsilosis*. The most prevalent Brg1 motif is listed as the representative motif. This representative motif shares high similarity between *C. albicans*, *C. dubliniensis*, and *C. parapsilosis*, but is absent in *C. tropicalis*. Instead, the representative Brg1 motif in *C. tropicalis* is identified in 3.7% of *C. albicans* Brg1 peaks (ALLR=0.852). Highly similar alternate motifs for Brg1 are listed in Figure 2.9. This Brg1 alternate motif is the sequence recognized by Ndt80. The Bcr1 and Brg1 motifs identified in this study are different from those identified by Nobile et al. (2012) and Mancera et al. (2015) (Nobile et al. (2012); Mancera et al. (2021)). Mancera et al. (2021) did not identify a motif for Bcr1 and the Brg1 motif they reported was similar to the Efg1 motif, likely due to shared target genes between Brg1 and Efg1. Both the primary and alternate Efg1 motifs are conserved across the four species. The primary Efg1 motif has higher entropy (i.e. decreased IC) in most positions in *C. dubliniensis* and *C. tropicalis* relative to *C. albicans* and *C. parapsilosis*. The alternate Efg1 motif, on the other hand, is more conserved (min(pairwise ALLR) = 0.949 and (max(pairwise ALLR) = 1.16) across the four species than the primary Efg1 motif (min(pairwise ALLR) = 0.092 and max(pairwise ALLR) = 0.717). Highly conserved primary (min(pairwise ALLR) = 0.953 and max(pairwise ALLR) = 1.124) and alternate motifs (min(pairwise ALLR) = 1.03 and max(pairwise ALLR) = 1.077) were also identified for Ndt80. The primary Rob1 motif in *C. albicans* and *C.tropicalis* is similar (ALLR=0.905), but the *C. dubliniensis* motif is distinct. The alternate motifs for Rob1 are species-specific, and are not similar to the primary or secondary motifs in other species. Notably, Tec1 motifs are the most dissimilar between species. The primary Tec1 motifs of *C. dubliniensis* and *C. tropicalis* are similar to the *C. albicans* alternate Tec1 motif, with ALLR of 0.57 and 0.333, respectively. Tec1 motifs in *C. parapsilosis* are unique to this species. Additionally, 46.6% of Tec1 binding loci in *C. parapsilosis* contain the Brg1 primary motif and 18.6% contain the Ndt80 primary motif.

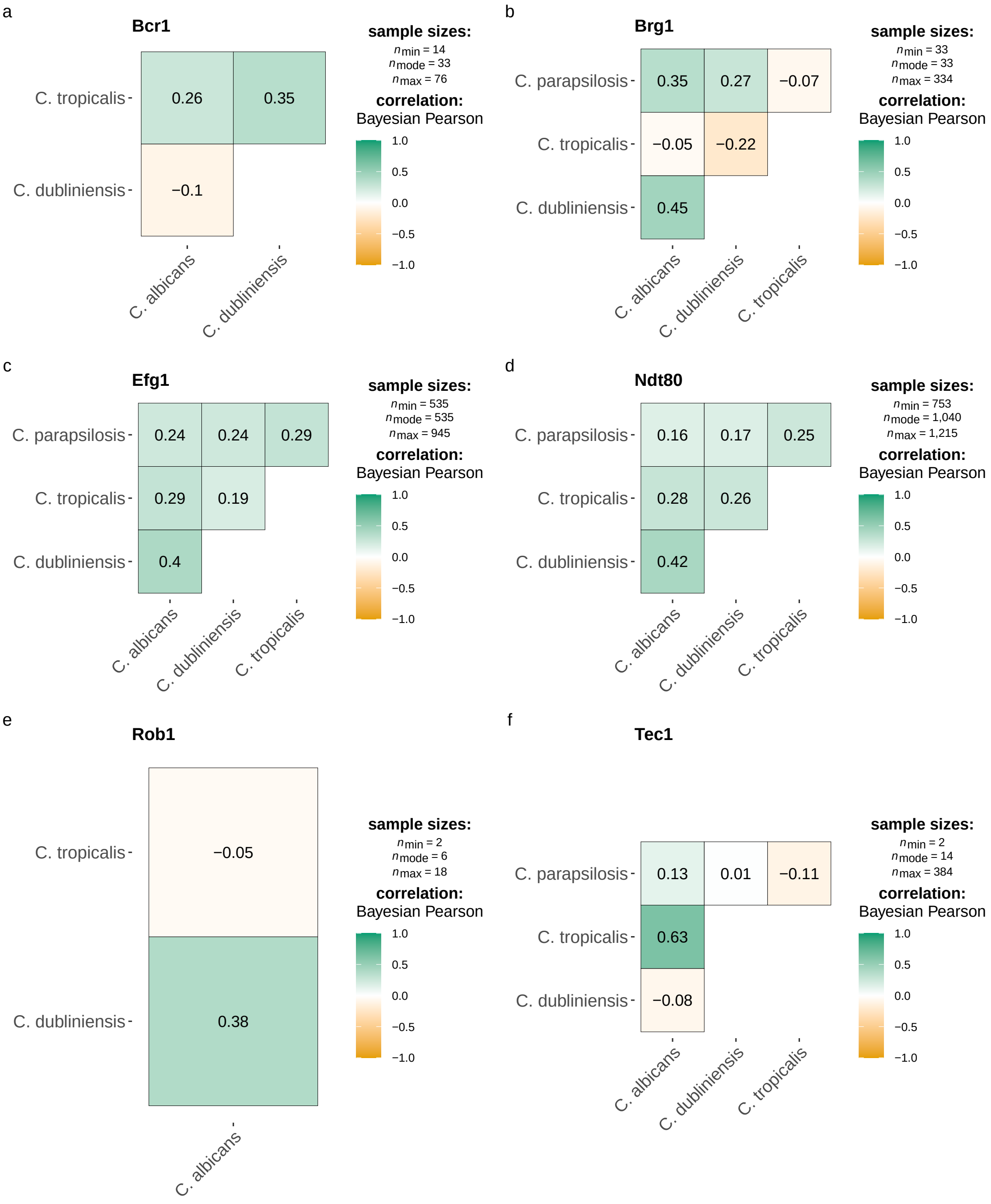


Figure 2.7: **Cumulative binding strength of Rob1 and Tec1 are least correlated in the diploid CTG species.**
Correlation of cumulative binding between species. Cumulative binding is the sum of binding strengths in the *cis*-regulatory regions of target genes. The correlation between binding strengths is shown as the Bayesian Pearson correlation coefficient. Efg1 and Ndt80 cumulative binding strengths are correlated in diploid CTG species, whereas Rob1 and Tec1 cumulative binding strengths are least correlated.

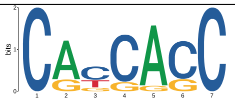
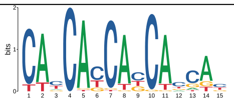
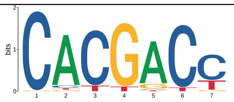
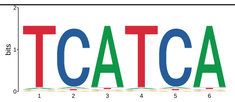
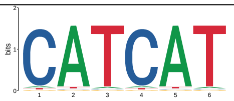
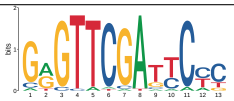
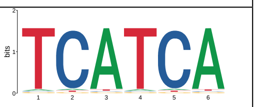
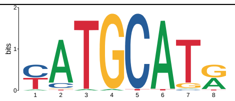
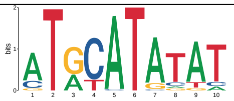
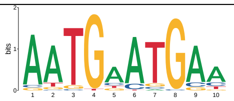
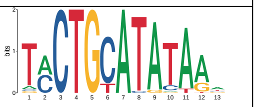
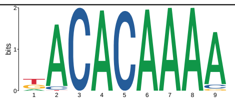
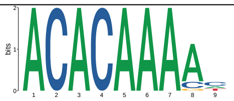
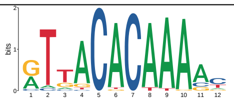
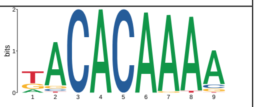
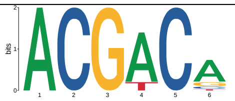
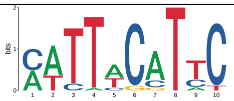
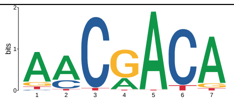
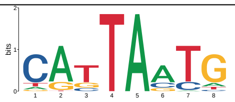
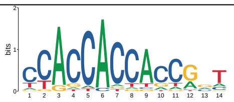
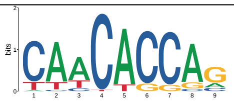
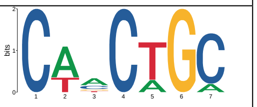
TF	*C. albicans*	*C. dubliniensis*	*C. tropicalis*	*C. parapsilosis*
Bcr1	 Total IC = 9.93 Sites = 38.5%	 Total IC = 15.82 Sites = 38.5%	 Total IC = 9.99 Sites = 28.2%	-
Brg1	 Total IC = 9.29 Sites = 33.9%	 Total IC = 9.25 Sites = 44.4%	 Total IC = 9.99 Sites = 30.5%	 Total IC = 9.29 Sites = 48.2%
Efg1	 Total IC = 10.75 Sites = 15.8%	 Total IC = 13.45 Sites = 22.2%	 Total IC = 12.19 Sites = 38.7%	 Total IC = 18.75 Sites = 7.9%
Ndt80	 Total IC = 15.12 Sites = 58.9%	 Total IC = 15.34 Sites = 43.4%	 Total IC = 15.51 Sites = 16.3%	 Total IC = 14.35 Sites = 78.6%
Rob1	 Total IC = 10.14 Sites = 42.1%	 Total IC = 12.89 Sites = 67.5%	 Total IC = 9.21 Sites = 52.3%	-
Tec1	 Total IC = 8.88 Sites = 42%	 Total IC =13.19 Sites = 45.7%	 Total IC = 11.15 Sites = 29.2%	 Total IC = 10.03 Sites = 37.6%

Figure 2.8: **Primary motifs of five master regulators are conserved across diploid CTG species.**
Primary motifs representing binding preference of TFs were identified for the master regulators in *C. albicans*, *C. dubliniensis*, *C. tropicalis* and *C. parapsilosis* using STREME (Bailey; 2021). The primary motifs of Bcr1, Brg1, Efg1 and Ndt80 are similar across diploid CTG species. Rob1 primary motif in *C. dubliniensis* is different from *C. albicans* and *C. tropicalis*. Tec1 primary binding motifs are distinct across the diploid CTG species. The prevalence of these primary motifs of TFs in their corresponding binding sites are shown along with total IC of the motif.

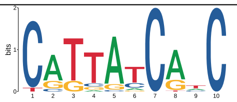
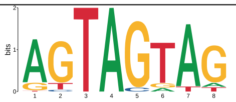
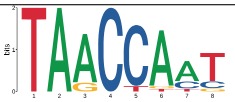
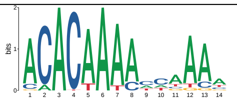
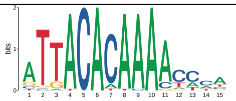
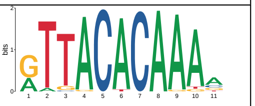
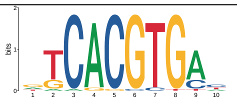
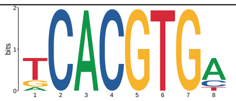
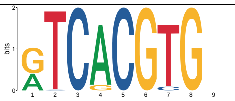
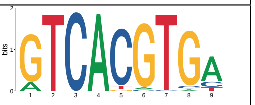
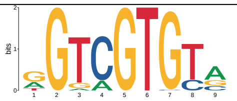
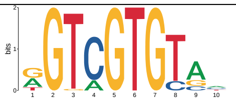
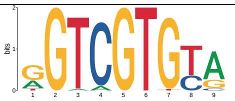
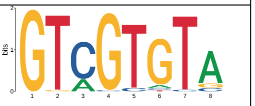
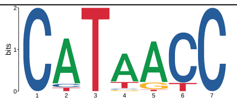
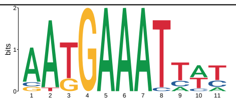
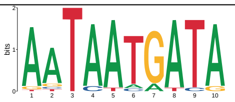
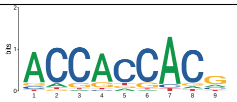
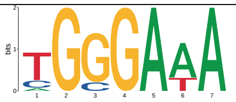
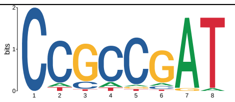
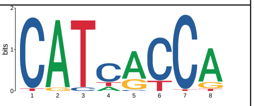
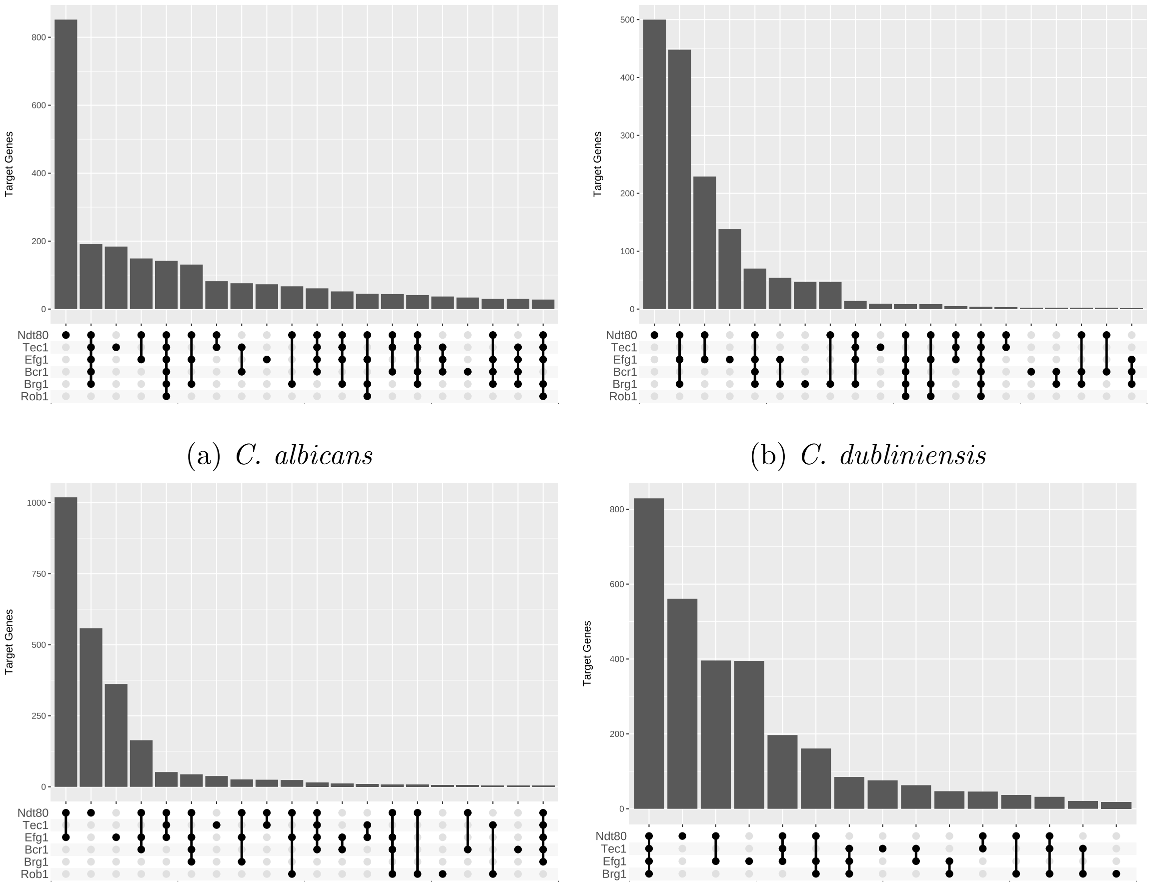
TF	*C. albicans*	*C. dubliniensis*	*C. tropicalis*	*C. parapsilosis*
Bcr1	 Total IC = 11.81 Sites = 20.9%	 Total IC = 12.01 Sites = 44.4%	 Total IC = 12.07 Sites = 23.7%	-
Brg1	 Total IC = 16.7 Sites = 17.6%	 Total IC = 16.36 Sites = 5.4%	-	 Total IC = 16.98 Sites = 17.6%
Efg1	 Total IC = 12.85 Sites = 6.4%	 Total IC = 13.32 Sites = 7.5%	 Total IC = 14.11 Sites = 10.8%	 Total IC = 14.2 Sites = 6.3%
Ndt80	 Total IC = 12.66 Sites = 15.7%	 Total IC = 13.84 Sites = 10.8%	 Total IC = 13.79 Sites = 18.4%	 Total IC = 12.47 Sites = 17.6%
Rob1	 Total IC = 10.77 Sites = 40.9%	 Total IC = 15.55 Sites = 47.5%	 Total IC = 15.69 Sites = 25.2%	-
Tec1	 Total IC = 8.36 Sites = 38.5%	 Total IC = 11.46 Sites = 43.5%	 Total IC = 11.19 Sites = 7.3%	 Total IC = 10.97 Sites = 28%

Figure 2.9: **Alternate motifs of Brg1, Efg1 and Ndt80 are conserved across diploid CTG species.**
Alternate motifs of TFs were identified for the master regulators in *C. albicans*, *C. dubliniensis*, *C. tropicalis* and *C. parapsilosis* using STREME (Bailey; 2021). These are second most prevalent motifs, following the primary motifs, in binding loci of the TFs. The alternate motifs of Brg1, Efg1 and Ndt80 are similar across diploid CTG species. The prevalence of these alternate motifs of TFs in their corresponding binding sites are shown along with total IC of the motif.



(a) *C. albicans*

(b) *C. dubliniensis*

(c) *C. tropicalis*

(d) *C. parapsilosis*

Figure 2.10: **Biofilm network structure has diverged between *C. albicans*, *C. dubliniensis*, *C. tropicalis* and *C. parapsilosis*.**
Number of target genes regulated by TF sets. The x-axis shows the set categories of target genes regulated by one or more TFs. The y-axis shows the number of target genes that belong to each category. Most of the target genes in the biofilm regulatory network in *C. albicans* and *C. dubliniensis* are regulated independently by Ndt80. In *C. dubliniensis* this is followed by the target genes regulated by three TFs, Ndt80, Efg1 and Brg1. Most of the target genes in the biofilm regulatory network in *C. tropicalis* are regulated by Ndt80 and Efg1 together, followed by Ndt80 and Efg1 independently. Most of the target genes in the biofilm regulatory network in *C. parapsilosis* are regulated by all four TFs, Brg1, Efg1, Ndt80 and Tec1.

2.3.5 Biofilm network structure has diverged across diploid CTG *Candida* species

In addition to the targets in the biofilm network, the network structure is also different in the four species (Figure 2.10). The number of TFs regulating each target gene is shown for *C. albicans* in Figure 2.10a, for *C. dubliniensis* in Figure 2.10b, for *C. tropicalis* in Figure 2.10c and for *C. parapsilosis* in Figure 2.10d. The biofilm network consists of 2633 genes in *C. albicans*, 1602 genes in *C. dubliniensis*, 2422 genes in *C. tropicalis* and 2964 genes in *C. parapsilosis*. Both the size and the structure of the biofilm regulatory network has diverged between these species. In all four species, in addition to regulating genes concertedly with other TFs, Ndt80 also regulates a large number of genes independently. Additionally, Efg1 and Ndt80 share the most targets among pairs of TFs in all the four species, showing that the interactions between Ndt80 and Efg1 are maintained across species. In *C. albicans*, following the genes uniquely regulated by Ndt80, the second frequent category of target genes is those regulated by five TFs, Bcr1, Brg1, Efg1, Ndt80 and Tec1. This shows that the biofilm network is highly connected in *C. albicans*. Rob1 does not independently regulate any genes in *C. albicans* or *C. dubliniensis*, suggesting that Rob1 is the last biofilm regulator to be integrated into this network. In *C. dubliniensis*, Ndt80, Efg1 and Brg1 regulate approximately 450 target genes. When comparing this category in *C. tropicalis* and *C. parapsilosis*, it is possible that *C. tropicalis* and *C. parapsilosis* lost regulatory interactions shared by these three TFs. Another, more parsimonious explanation is the increase in targets regulated by these three TFs in *C. dubliniensis* and *C. albicans*. In *C. tropicalis*, most of the target genes are regulated by Ndt80 and Efg1 together or independently, followed by Bcr1. This shows that these three regulators are more closely integrated in the *C. tropicalis* biofilm network compared to Brg1, Rob1 and Tec1. In *C. parapsilosis*, most of the genes in the network are regulated by all four TFs Brg1, Efg1, Ndt80 and Tec1. This is a unique aspect of the *C. parapsilosis* biofilm network, given that it is also the network with the most number of target genes. Another unique aspect of the *C. parapsilosis* biofilm network is the number of target genes regulated independently by Tec1. Since, Tec1 is not required for biofilm formation in this species, it is likely that these targets are involved in processes other than biofilm development.

We next wanted to quantify the difference in network structure between the four diploid CTG *Candida* species. We used the Jensen-Shannon Divergence (JSD) to measure differences in number of TFs regulating the same target gene (i.e. in-degree of the target gene in the biofilm network). The JSD between *C. albicans* and *C. dubliniensis* is 0.082 bits and the biofilm network structure is significantly different in these two species. Similarly, the JSD between the *C. albicans* and *C. tropicalis* biofilm networks is 0.121 bits and is significantly different in these two species. To compare in-degree distributions between different networks using JSD we need the same number of regulators. *C. parapsilosis* was not included in this comparison since we identified targets of only four regulators. Hence, we visualized the distributions of node in-degree across species (Figure 2.11), and we can see that *C. albicans* has more targets regulated by five or more TFs. A large number of target genes in *C. parapsilosis* are regulated by four TFs. This shows that the biofilm network is more interconnected in *C. albicans* and *C. parapsilosis*.

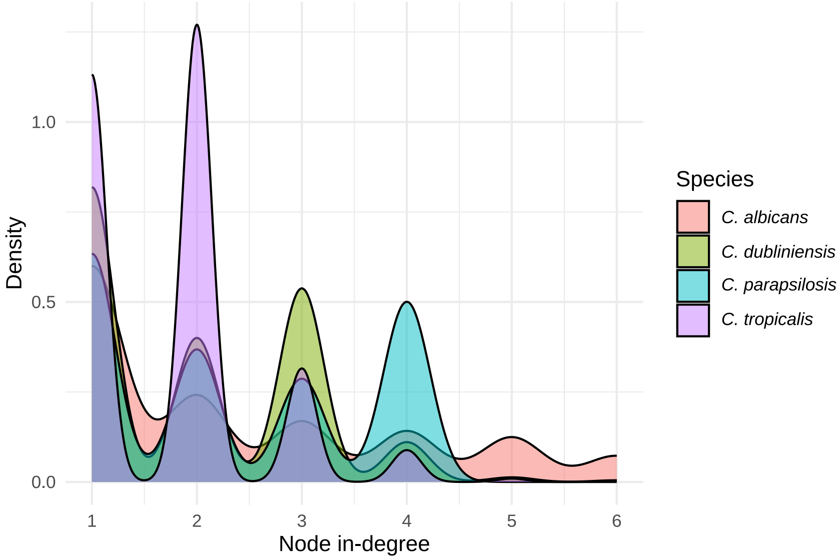


Figure 2.11: **Biofilm network structure varies in diploid CTG species.** Distribution of node in-degree in biofilm network across species. The x-axis denotes the node number of TFs regulating the target and the y-axis is probability density of target genes regulated by the corresponding number of TFs on the x-axis. *C. albicans* has more targets regulated by five or more TFs.

2.3.6 *C. albicans* target genes are enriched for host-fungal interactions

Since the TFs regulate different genes in the four diploid CTG *Candida* species, we wanted to determine whether the TFs regulate different functions in these species and whether they have any species-specific roles. We performed a gene-set functional enrichment analysis as described in subsection 2.5.4, using the TF targets in each species as the gene lists and all genes in the genome as the background. The functional enrichment results of selected GO terms for Bcr1 is shown in Figure 2.12, Brg1 in Figure 2.13, Efg1 in Figure 2.14, Ndt80 in Figure 2.15, and Tec1 in Figure 2.16. Rob1 is not included since biological processes (BP) functional categories were not enriched in *C. dubliniensis* and *C. tropicalis*. All of the TFs have acquired targets with new roles in *C. albicans*, regulating additional functions compared to *C. dubliniensis*, *C. tropicalis* and *C. parapsilosis*. Bcr1 regulates filamentation in *C. albicans* and

C. dubliniensis, and additionally regulates cell adhesion, interspecies interaction, entry into host and adhesion of symbiont to host, exclusively in *C. albicans.* Brg1 regulates hyphal cell wall formation and biofilm formation on inanimate substrate across multiple species. Brg1 also regulates some functions exclusively in *C. albicans*, including adhesion of symbiont to host, development of symbiont in host and entry into host. Efg1 regulates interspecies interaction in all four diploid CTG *Candida* species. In *C. tropicalis*, Efg1 regulates pseudohyphal growth, as well as response to stresses, such as heat and glucose starvation. Similar to other TFs, Efg1 regulates cell-cell adhesion, adhesion of symbiont to host, development of symbiont in host, entry into host and induction by symbiont of host defense response, exclusively in *C. albicans.* Ndt80, similar to Efg1 regulates essential functions in *C. tropicalis* such as signal transduction, regulation of mitotic cell cycle and response to heat. In *C. albicans*, Ndt80 also regulates adhesion of symbiont to host and entry into host. Tec1 regulates essential functions in *C. parapsilosis.* These functions include protein folding, regulating chaperone functions and response to glucose starvation. In *C. albicans*, Tec1 along with other TFs regulates adhesion to host and entry into host. Although hyphal formation and biofilm formation is enriched in multiple species, fungal-host relationship is regulated uniquely in *C.albicans.* The integration of these functions into the biofilm network, especially by multiple TFs, likely contribute to robust biofilm formation in *C. albicans.*

2.4 Discussion

In this study, we assessed the conservation of biofilm regulators and target genes across species. To do so, we carefully verified the orthologous relationships inferred for the TFs. OrthoFinder was not able to identify an ortholog of *EFG1* in *C.tropicalis*, but this is due to an assembly gap in the *C. tropicalis* genome (Mancera et al. (2015)). Similarly, OrthoFinder was unable to find the ortholog of *FLO8* in *S. cerevisiae.* This is possibly due to a missense mutation in the *S. cerevisae* reference strain S288C (Liu et al. (1996)). Hence, the choice of reference and high quality assembly are important for the interpretation of the results. Another example is *GAL4*, where *GAL4* is present in *S. cerevisiae*, and shares a homologous DNA-binding domain with the *C. albicans GAL4* (Askew et al. (2009)). However, *S. cerevisiae GAL4* belongs to a different orthogroup and homologs of this gene are present in most of the CTG *Candida* species. Even though the *S. cerevisiae GAL4* and *C. albicans GAL4* have similar binding domains and likely shared a common ancestor, the are not clustered in the same orthogroup, likely due to the divergence in their activation domain (Martchenko et al. (2007)).

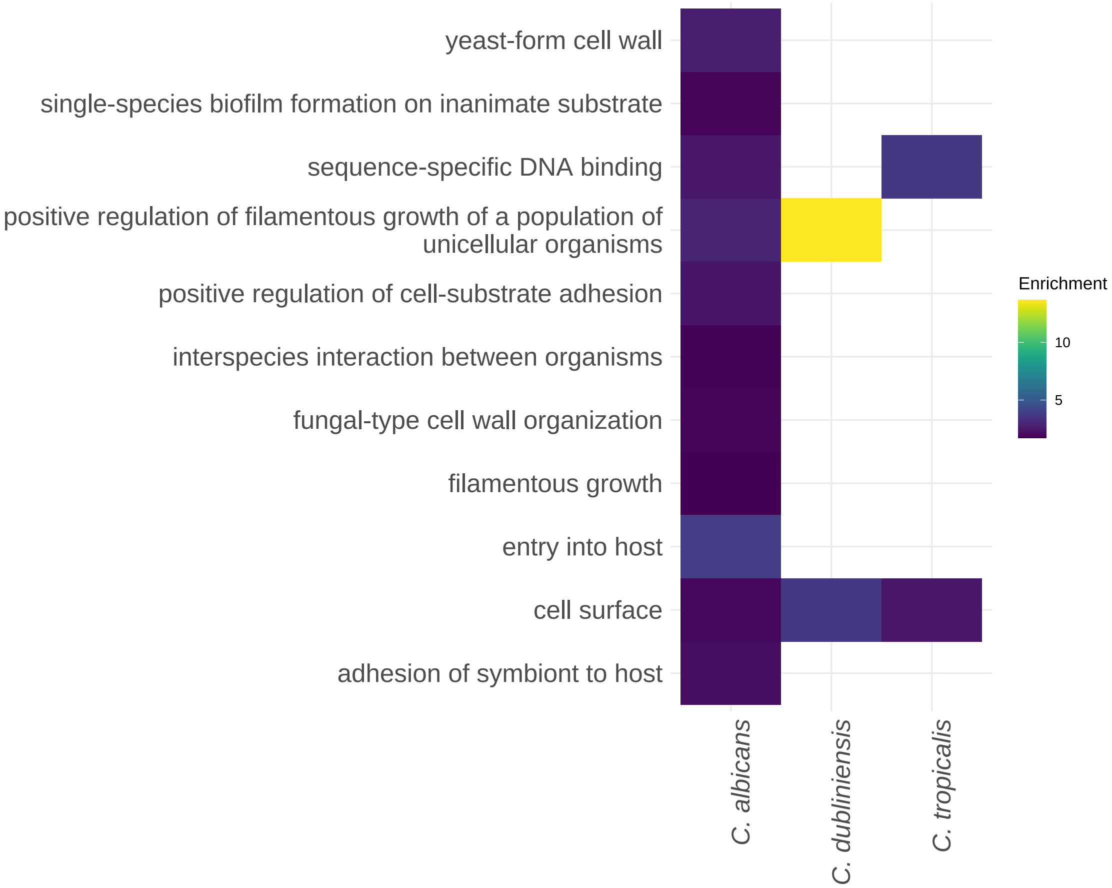


Figure 2.12: **Bcr1 regulates entry into host and interspecies interaction in *C. albicans*.**
A subset of GO terms enriched in target genes of Bcr1 in *C. albicans*, *C. dubliniensis* and *C. tropicalis*. The y-axis shows the biological processes enriched in at least one of the above species. The color in the heatmap shows the enrichment score which is the proportion of genes annotated with the corresponding GO term in the target gene list over proportion of genes annotated with the GO term in the background gene set. Bcr1 regulates interspecies interaction, entry into host and adhesion of symbiont to host, exclusively in *C. albicans*.

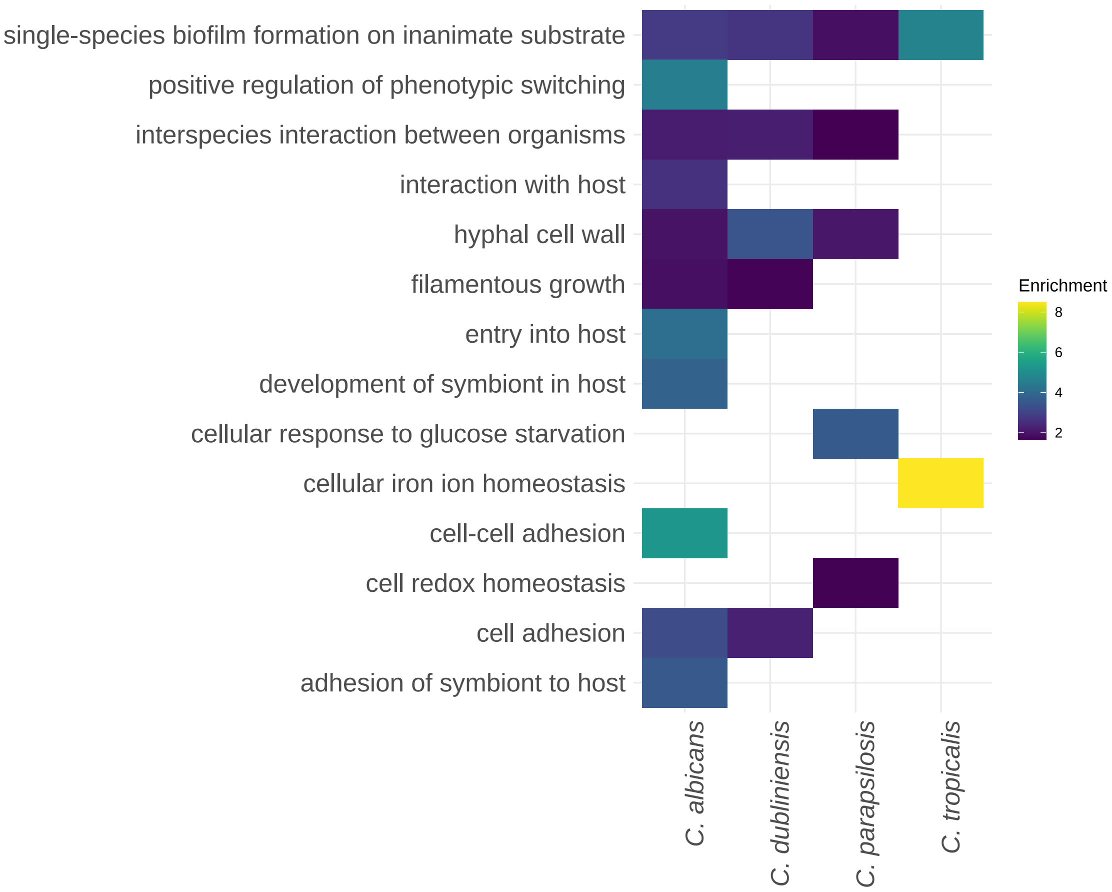


Figure 2.13: **Brg1 regulates host-pathogen functions in *C. albicans*.**
A subset of BP GO terms enriched in target genes of Brg1 in *C. albicans*, *C. dubliniensis*,*C. parapsilosis* and *C. tropicalis*. The y-axis shows the biological processes enriched in at least one of the above species. The color in the heatmap shows the enrichment score which is the proportion of genes annotated with the corresponding GO term in the target gene list over proportion of genes annotated with the GO term in the background gene set. Brg1 regulates biofilm formation and interspecies interaction in multiple species. Whereas, Brg1 regulates adhesion of symbiont to host, development of symbiont in host and entry into host exclusively in *C. albicans*.

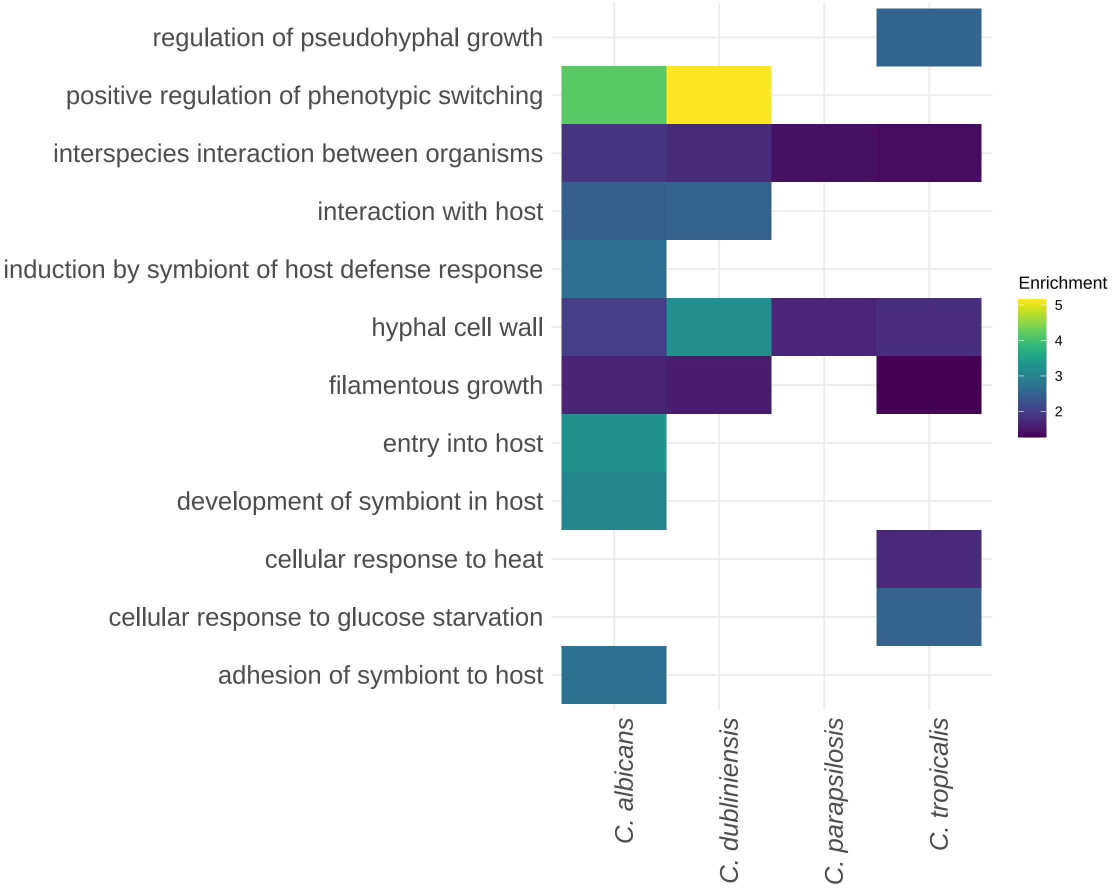


Figure 2.14: **Efg1 regulates interspecies interaction in all four diploid CTG *Candida* species.**

A subset of BP GO terms enriched in target genes of Efg1 in *C. albicans*, *C. dubliniensis*, *C. parapsilosis* and *C. tropicalis*. The y-axis shows the biological processes enriched in at least one of the above species. The color in the heatmap shows the enrichment score which is the proportion of genes annotated with the corresponding GO term in the target gene list over proportion of genes annotated with the GO term in the background gene set. Efg1 regulates interspecies interaction in all four diploid CTG *Candida* species. Efg1 regulates cell-cell adhesion, adhesion of symbiont to host, development of symbiont in host, entry into host and induction by symbiont of host defense response, exclusively in *C. albicans*.

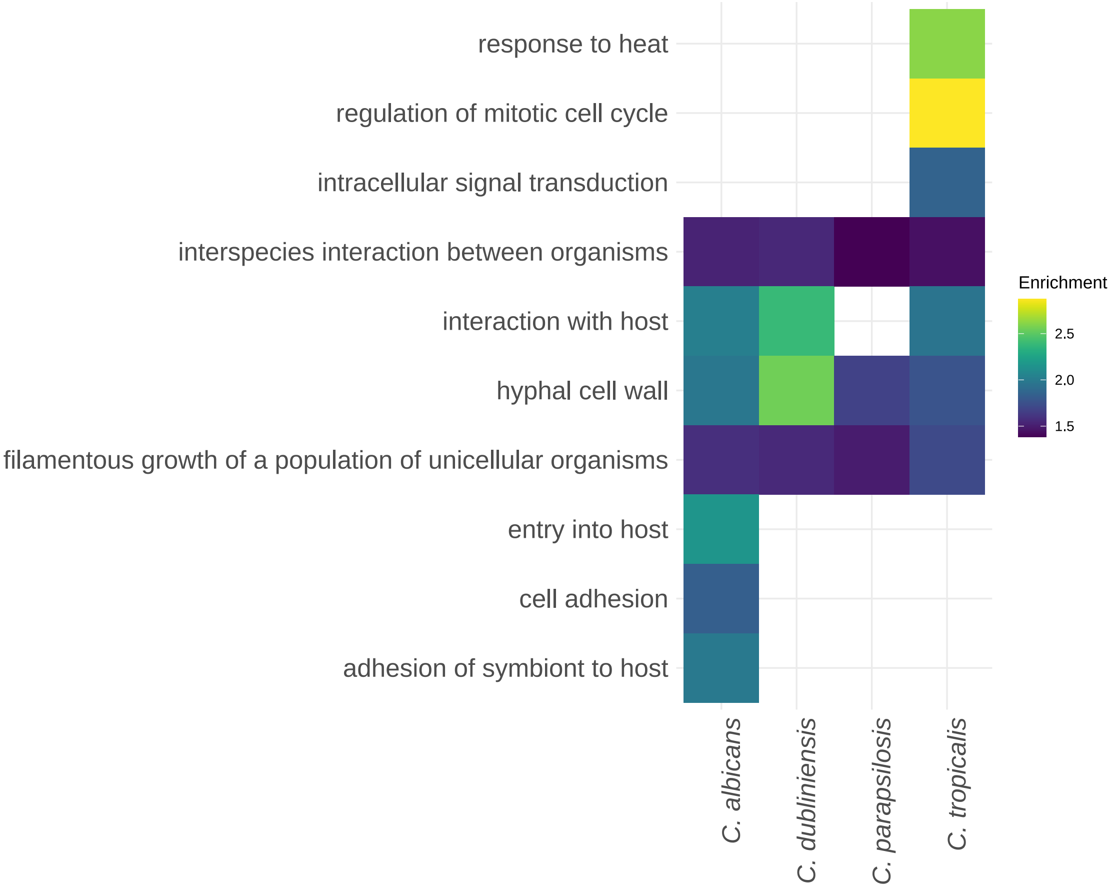


Figure 2.15: **Ndt80 regulates host-pathogen functions in *C. albicans*.** A subset of BP GO terms enriched in target genes of Ndt80 in *C. albicans*, *C. dubliniensis* and *C. tropicalis* and *C. parapsilosis*. The y-axis shows the biological processes enriched in at least one of the above species. The color in the heatmap shows the enrichment score which is the proportion of genes annotated with the corresponding GO term in the target gene list over proportion of genes annotated with the GO term in the background gene set. Ndt80 regulates essential functions in *C. tropicalis*. In *C. albicans*, Ndt80 regulates adhesion of symbiont to host and entry into host.

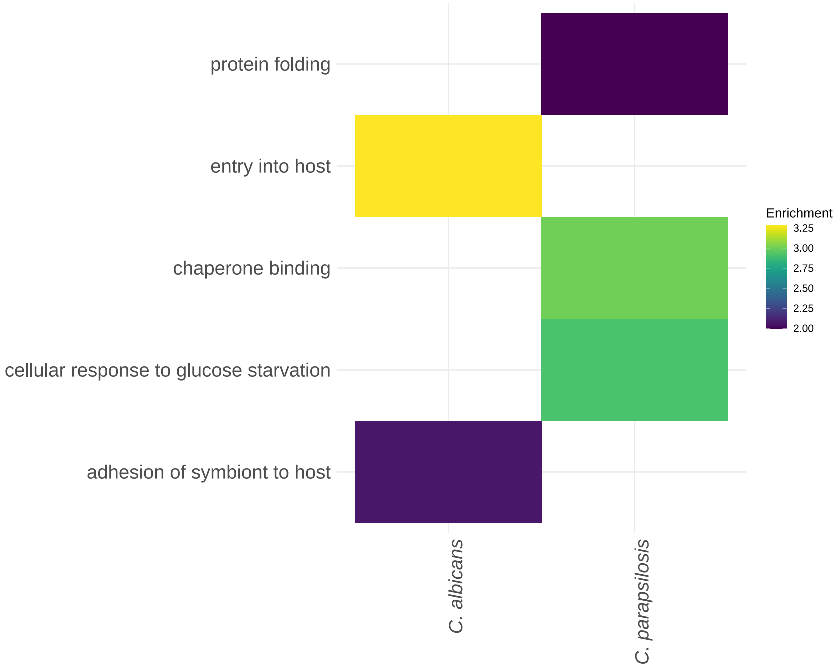
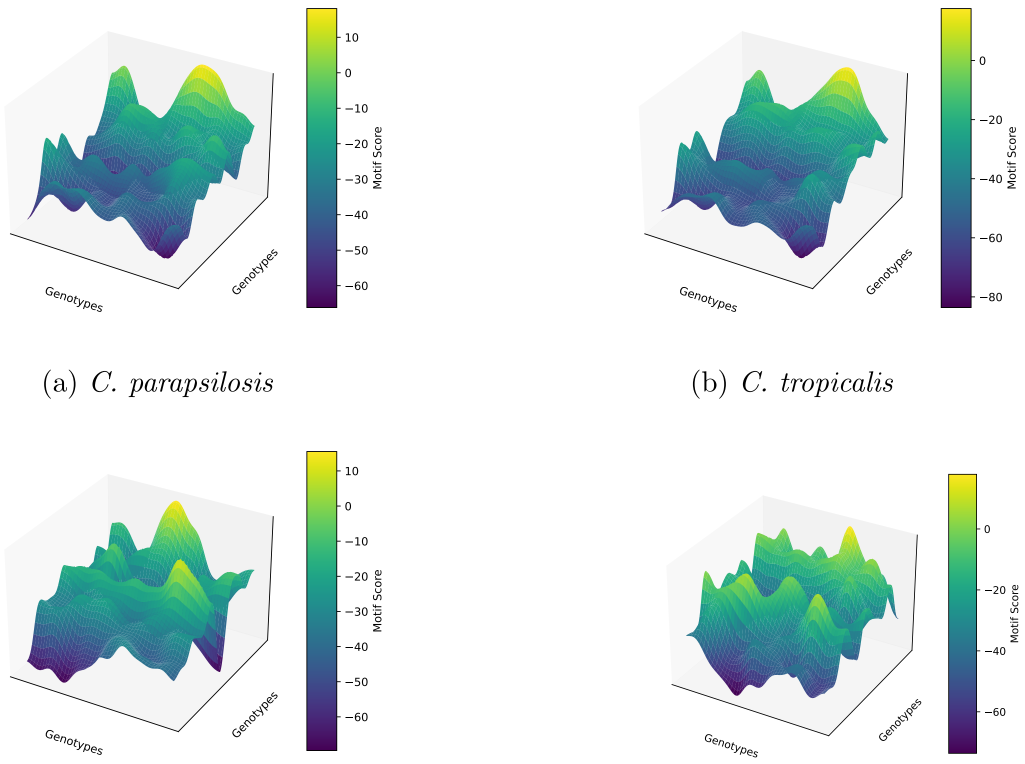


Figure 2.16: **Tec1 regulates regulates adhesion to host and entry into host in *C. albicans*.**
A subset of BP GO terms enriched in target genes of Tec1 in *C. albicans* and *C. parapsilosis*. No BPs were enriched in *C. tropicalis*. The y-axis shows the biological processes enriched in at least one of the above species. The color in the heatmap shows the enrichment score which is the proportion of genes annotated with the corresponding GO term in the target gene list over proportion of genes annotated with the GO term in the background gene set. Tec1 regulates essential functions in *C. parapsilosis*. In *C. albicans*, Tec1 along with other TFs regulates adhesion to host and entry into host.

We found that CTG and non-CTG species have diverged considerably in their TFs and target genes and even well conserved TFs largely regulate "younger" genes (genes unique to the CTG species). Divergence in TF roles in non-CTG species is likely driven by multiple TF duplication events in this clade. But variation in CTG species are driven predominantly by *cis*-acting variations. This is supported by the conservation of both TFs, target genes and binding preferences of TFs. Binding motifs for Ndt80 are conserved across large evolutionary distances, where this TF has the same motif preference in *S. cerevisiae* (Nocedal et al. (2017)). Tec1 binding preference is distinct in *C. parapsilosis*. Interestingly, *C. parapsilosis* does not require Brg1 and Tec1 for biofilm formation. In *C. parapsilosis*, Bcr1 motifs were found in Brg1 binding sites and Ndt80 and Bcr1 motifs were observed in Tec1 binding sites. Hence, it is likely that targets of Tec1 and Brg1 are regulated by other TFs in *C. parapsilosis*. Also of note is that even when the binding preference of a TF is similar across species, the mutational robustness of the motif likely varies. For instance, Figure 2.17 shows the mutational robustness of the Efg1 motif in *C. albicans*, *C. dubliniensis*, *C. tropicalis* and *C. parapsilosis*. The construction of the landscape is detailed in subsection 2.5.3. The width of the global peak (highlighted in yellow) is indicative of mutational tolerance for the motif (Aguilar-Rodríguez et al. (2017); Payne and Wagner (2014)). The Efg1 motif peak is broader in *C. parapsilosis* compared to the *C. albicans*, indicating that mutations in *C. albicans* are more likely to result in a loss of function, than in *C. parapsilosis*. The mutational landscape should be further validated by comparing binding strengths in *cis*-regulatory regions in addition to motif strength.



(a) *C. parapsilosis*

(b) *C. tropicalis*

(c) *C. dubliniensis*

(d) *C. albicans*

Figure 2.17: **Mutational landscape of Efg1 motif.**
The landscape of motif scores. The x-axis and y-axis indicate all possible k-mers, where k is the motif length and the z-axis denotes the motif score calculated from the PPM of the motif. The breadth of the global peak corresponds to mutational robustness. Mutations in *C. albicans* are more likely to result in a loss of function, than in *C. parapsilosis*.

The biofilm network structure has significantly changed between *C. albicans* and its close relatives. Even though Tec1 and Brg1 are not required for biofilm formation in *C. parapsilosis*, these regulators are well integrated with the other biofilm regulators in this species. This suggests that Brg1 and Tec1 likely function as redundant regulators in the biofilm network as opposed to cooperatively regulating target genes along with Ndt80 and Efg1. This redundancy could have resulted in divergence of target genes regulated by these TFs, making them essential for biofilm formation in *C. tropicalis*, *C. dubliniensis* and *C. albicans*.

We also found that the biofilm regulatory network in *C. albicans* has acquired new functions, specifically pertaining to symbiont interaction with host. One BP category that was enriched in targets of all TFs in *C. albicans* is adhesion of symbiont to host. The enrichment of this functional category is driven by GPI-anchored cell wall adhesins, such as Eap1. The gene family encoding GPI-anchor proteins has expanded in *C. albicans* relative to non-CTG *Candida* species (Nather and Munro (2008)). Additionally, adhesins have been shown to rescue the ability to form biofilms in the absence of upstream regulators, emphasizing the integral role of adhesion in biofilm development (Nobile et al. (2006)). Another noteworthy enriched category is development of symbiont in host. This category is annotated for genes whose roles involve progression of an organism from initial condition, either through filamentation or secretion of proteases. The enrichment of this functional category is driven by transcription factors involved in filamentation as well as regulators of mating and pheromone-simulated biofilm formation (such as Cph1) (Lin et al. (2013)).

In summary, we found that duplication of TFs and gene loss are not responsible for differences in biofilm formation in the CTG species. Binding preferences of TFs are also largely conserved, but the network structure varies considerably with the *C. albicans* biofilm network being highly interconnected compared to its close relatives. This rewiring has also resulted in the recruitment of new functions to the network that enable *C. albicans* to tolerate and leverage changes in the host environment. One caveat to note is that not all interspecific binding events predicted using ChIP-Seq data have functional consequences. Other factors such as co-binding of other regulators, chromatin accessibility and local genomic context also play roles in determining functional consequence of binding (Jiang and Mortazavi; 2018). A recent study comparing *S. cerevisiae* and *C. albicans* regulators involved in iron-uptake shows that nonfunctional binding is pervasive and specifically advantageous in adaptation to new environmental conditions (Hsu et al.; 2021). Hence, divergence in non-functional binding are also indicative of network rewiring.

2.5 Methods

2.5.1 Data acquisition

The most complete genome assembly and annotation (in FASTA format) was retrieved (between January 26th - February 1st 2023) for the yeast species in this study from GenBank, RefSeq and the National Center for Biotechnology Information (NCBI). The assembly accession numbers along with assembly completeness, genome size, and number of annotated protein-coding genes are listed in Table 2.1. The genome assembly of *C. pelliculosa* was fragmented into 46 scaffolds, so this species was removed from further analyses. The gene expression data, identifying genes differentially expressed in biofilm relative to planktonic growth conditions, as well as ChIP-ChIP binding of biofilm regulators in *C. albicans* was retrieved from Nobile et al. (2012). Binding of biofilm regulators in *C. albicans*, *C. dubliniensis*, *C. tropicalis* and *C. parapsilosis* identified using ChIP-Seq experiments were retrieved from Mancera et al. (2021).

2.5.2 Identifying orthologous genes and regions

Orthologs and paralogs of *C. albicans* genes were identified across species using OrthoFinder (version 2.5.4) (Emms and Kelly (2015, 2019)). OrthoFinder identifies orthogroups by computing similarity scores between proteins using an all-vs-all BLAST search. The similarity scores are normalized for gene lengths and phylogenetic distance. These normalized scores are then used to construct an orthogroup graph, where the nodes are genes and edge weights are given by the normalized scores. These normalized scores are then used to identify orthogroups using Markov Clustering (MCL) (Van Dongen (2008)). OrthoFinder was run with parameters `-S diamond` to identify sequence similarity using DIAMOND (Buchfink et al. (2015)) and `-M msa` to perform multiple sequence alignment (MSA) of orthogroups with MAFFT (Katoh and Standley (2013)) and infer gene trees from MSA using FastTree (Price et al. (2010)). OrthoFinder uses a set of gene trees that include genes from all species, including multi-copy gene families to infer the unrooted species tree using the Species Tree from All Genes (STAG) algorithm (Emms and Kelly (2018)). OrthoFinder then uses the Species Tree Root Inference from Duplication Events (STRIDE) algorithm to identify gene duplication events in unrooted orthogroup trees and uses these events to infer the species root (Emms and Kelly (2017)). The rooted species tree is then used to root the gene trees and infer orthologs and gene duplication events using duplication-loss-coalescence (DLC) analysis. DLC is used to infer the most parsimonious history of a gene family, accounting for gene duplication, gene loss and incomplete lineage sorting (Wu et al. (2014)).

2.5.3 ChIP-Seq data analysis

Publicly available ChIP-Seq data was retrieved from Mancera et al. (2021) to identify binding targets of the six master biofilm regulators, Bcr1, Brg1, Efg1, Ndt80, Rob1 and Tec1 in *C. albicans*, *C. dubliniensis* and *C. tropicalis.* ChIP-Seq data for four biofilm regulators (excluding Rob1 and Bcr1) was available for *C. parapsilosis.* The analysis of single-end ChIP-Seq data is described in detail in chapter 4. Briefly, sequences were aligned to the reference genomes of the respective *Candida* species using Bowtie2 (version 2.2.5) (Langmead and Salzberg (2012)). Peaks were called using MACS2 (version 2.2.7.1) (Zhang et al. (2008)). A custom python script was used to annotate the peaks and retrieve sequences to identify binding motifs using STREME from the MEME-suite (version 5.5.2) (Bailey (2021)) and the memes package (version 1.2.5) in R (Nystrom and McKay (2021)). Similarity between motifs was estimated using the average log-likelihood ratio (ALLR) (Wang and Stormo (2003)). A mutational landscape was also visualized using a custom python (version 3.10) script. All possible k-mers were obtained, where k is the length of the motif. If $k > 10$, then ends of the motifs were trimmed, to obtain a minimum bit score of one at the start and end positions of the motif. Motif trimming was performed using the universalmotif package (version 1.12.4) in R. The position probability matrix (PPM) was retrieved for the motif, which indicates the probability of observing a nucleotide n at position i. The motif score was then calculated as shown below:

$$\text{Motif Score} = \sum_{i}^{k} \log_2 \frac{P_n}{B_n}$$

where, P_n is the probability of nucleotide n at position i of the k-mer and B_n is the background frequency of nucleotide n. The mutational landscape was visualized for the motif using the matplotlib package (version 3.6.2) in python, where all k-mers were ordered by their Hamming distances to each other and arranged in a grid, with the z-axis denoting the motif score.

2.5.4 Functional enrichment

Functions regulated by each TF in the four diploid CTG *Candida* species mentioned above, were inferred as follows. We obtained Gene Ontology (GO) annotations for the *C. albicans* protein-coding genes from the Candida Genome Database (CGD) (Skrzypek et al. (2016)). We then assigned GO annotation to *C. dubliniensis*, *C. tropicalis* and *C. parapsilosis* genes based on their orthologous relationships to *C. albicans* using OrthoFinder (Emms and Kelly (2015, 2019)). Orthologous gene(s) in *C. albicans* were identified for approximately 98% of genes in *C. dubliniensis*, 90% of genes in *C. tropicalis* and 93% of genes in *C. parapsilosis.* After obtaining the target genes for each TF, we performed gene list enrichment to identify enriched biological processes (BP), cellular components (CC), and molecular functions (MF).

Gene list enrichment analysis was performed using a hypergeometric test with the clusterProfiler package (version 4.2.2) in R (version 4.1.2) (Wu et al. (2021)). The GO categories annotated in target genes were compared against all annotated functions in *C. albicans* and the p-value denoting significance of enrichment was corrected for multiple testing using FDR (Benjamini and Hochberg; 1995).

2.5.5 Data visualization

Statistical tests, violin plots, and correlation matrix visualizations were performed using the ggstatsplot package (version 0.11.1) in R (Patil (2021)). Venn diagrams were created using the ggvenn package (version 0.1.10) in R. The functional enrichment heatmaps were created using ggplot2 package (version 3.4.2) in R. Phylogenetic trees were visualized using the Interactive Tree of Life (iTol) (Letunic and Bork (2021)). R version 4.1.2 was used for all R packages.

2.5.6 Statistics and statistical tests

Friedman test

The Friedman test (Friedman; 1937) is a non-parametric test for comparing differences between groups for dependent samples. The test assumes that the dependent variable is ordinal or continuous. The dependent variable need not be normally distributed. The test statistic is given as:

$$\chi^2(Q) = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

where, n is the number of samples, k is the number of groups, R_j is the sum of ranks for the group j.

Jensen-Shannon Divergence (JSD)

JSD measures the difference between two probability distributions (Nielsen (2019)). This metric was used to calculate the difference in degree distributions between networks. The in-degree of nodes in the network, i.e. the number of TFs regulating each gene was retrieved and the probability mass function (PMF) of the in-degree distribution was computed. The JSD of in-degree distribution of two networks are given as:

$$JSD(P||Q) = 0.5(KL(P||R) + KL(Q||R))$$

where, P and Q are the PMF of node in-degrees of network p and network q, respectively. $R = 0.5(P + Q)$ i.e. the midpoints of probability vectors P and Q and $KL(P||R)$ is the Kullback-Liebler Divergence (KLD) of P and R, computed as

$KL(P||R) = 0.5 \sum_{x \in X} P(x) \log_2(\frac{P(x)}{R(x)})$ (Kullback and Leibler (1951)). JSD is bounded by 1, i.e. $0 \leq JSD(P||Q) \leq 1$, where 0 indicates that P and Q are identical, whereas 1 indicates maximum divergence and shows that P and Q are completely dissimilar.

The confidence interval and p-value of the observed JSD was estimated using a permutation test by randomly sampling the in-degree of nodes for the two networks and calculating the JSD for the permuted networks. This sampling was repeated 1000 times and p-value was computed as the proportion of permuted networks having a JSD greater than the observed JSD.

Information Content (IC)

IC was calculated using the memes package (version 1.2.5) in R (Nystrom and McKay (2021)). IC shows the relative entropy at each position and is given using the formula below:

$$IC(i) = \sum_{n \in A,C,G,T} P(i) \times \log_2(\frac{P(i)}{B_i})$$

where, $P(i)$ is the probability of the nucleotide in position i and $B(i)$ is the background frequency of the nucleotide.

Average Log-likelihood Ratio (ALLR)

Representative motifs were selected for each TF in each species by comparing all motif hits using ALLR, to consolidate similar motif hits (Wang and Stormo (2003)). To compare two motifs, all possible alignments of the motifs were retrieved and the ALLR score was computed for all alignments and the best score was reported as the motif similarity metric. The ALLR score was calculated between a pair of columns in an aligned motif as given below:

$$ALLR = \frac{\sum_{n \in A,C,G,T} c_{nj} \ln \frac{f_{ni}}{P_n} + \sum_{n \in A,C,G,T} c_{ni} \ln \frac{f_{nj}}{P_n}}{\sum_{n \in A,C,G,T} c_{ni} + c_{nj}}$$

where, i and j are columns in the two alignment matrices. c_{ni} and c_{nj} are count vectors for nucleotide n and f_{ni} and f_{nj} are frequency vectors estimated from the observed counts plus some pseudocounts to reduce small sample biases. P_n denotes the background frequency of nucleotide n. Essentially, $\sum_{n \in A,C,G,T} c_{nj} \ln \frac{f_{ni}}{P_n}$ is the likelihood of the observed column j being generated by the distribution estimated from column i. Thus, ALLR measures the joint probability of observing the data generated by one distribution, given the likelihood ratio of the other distribution over background nucleotide frequencies (Wang and Stormo (2003)). Higher positive values of ALLR indicate high similarity and negative values indicate low similarity.

2.6 Bibliography

Aguilar-Rodríguez, J., Payne, J. L. and Wagner, A. (2017). A thousand empirical adaptive landscapes and their navigability, *Nature ecology & evolution* 1(2): 0045.

Askew, C., Sellam, A., Epp, E., Hogues, H., Mullick, A., Nantel, A. and Whiteway, M. (2009). Transcriptional regulation of carbohydrate metabolism in the human pathogen Candida albicans, *PLoS pathogens* 5(10): e1000612.

Bailey, T. L. (2021). STREME: accurate and versatile sequence motif discovery, *Bioinformatics* 37(18): 2834–2840.

Baker, C. R., Booth, L. N., Sorrells, T. R. and Johnson, A. D. (2012). Protein modularity, cooperative binding, and hybrid regulatory states underlie transcriptional network diversification, *Cell* 151(1): 80–95.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal statistical society: series B (Methodological)* 57(1): 289–300.

Buchfink, B., Xie, C. and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND, *Nature methods* 12(1): 59–60.

Byrne, K. P. and Wolfe, K. H. (2007). Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication, *Genetics* 175(3): 1341–1350.

Cavalheiro, M. and Teixeira, M. C. (2018). Candida biofilms: threats, challenges, and promising strategies, *Frontiers in medicine* 5: 28.

Cendejas-Bueno, E., Kolecka, A., Alastruey-Izquierdo, A., Theelen, B., Groenewald, M., Kostrzewa, M., Cuenca-Estrella, M., Gómez-López, A. and Boekhout, T. (2012). Reclassification of the Candida haemulonii complex as Candida haemulonii (C. haemulonii group I), c. duobushaemulonii sp. nov.(C. haemulonii group II), and C. haemulonii var. vulnera var. nov.: three multiresistant human pathogenic yeasts, *Journal of clinical microbiology* 50(11): 3641–3651.

Choudhury, B. I. and Whiteway, M. (2018). Evolutionary transition of GAL regulatory circuit from generalist to specialist function in ascomycetes, *Trends in microbiology* 26(8): 692–702.

Chow, N. A., Muñoz, J. F., Gade, L., Berkow, E. L., Li, X., Welsh, R. M., Forsberg, K., Lockhart, S. R., Adam, R., Alanio, A. et al. (2020). Tracing the evolutionary history and global expansion of Candida auris using population genomic analyses, *MBio* 11(2): e03364–19.

Cleary, I. A., Lazzell, A. L., Monteagudo, C., Thomas, D. P. and Saville, S. P. (2012). BRG1 and NRG1 form a novel feedback circuit regulating Candida albicans hypha formation and virulence, *Molecular microbiology* 85(3): 557–573.

Dalal, C. K., Zuleta, I. A., Mitchell, K. F., Andes, D. R., El-Samad, H. and Johnson, A. D. (2016). Transcriptional rewiring over evolutionary timescales changes quantitative and qualitative properties of gene expression, *Elife* 5: e18981.

Do, E., Cravener, M. V., Huang, M. Y., May, G., McManus, C. J. and Mitchell, A. P. (2022). Collaboration between antagonistic cell type regulators governs natural variation in the Candida albicans biofilm and hyphal gene expression network, *MBio* 13(5): e01937–22.

Edmond, M. B., Wallace, S. E., McClish, D. K., Pfaller, M. A., Jones, R. N. and Wenzel, R. P. (1999). Nosocomial bloodstream infections in United States hospitals: a three-year analysis, *Clinical infectious diseases* 29(2): 239–244.

Emms, D. and Kelly, S. (2018). STAG: species tree inference from all genes, *BioRxiv* p. 267914.

Emms, D. M. and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy, *Genome biology* 16(1): 1–14.

Emms, D. M. and Kelly, S. (2017). STRIDE: species tree root inference from gene duplication events, *Molecular biology and evolution* 34(12): 3267–3278.

Emms, D. M. and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics, *Genome biology* 20: 1–14.

Fox, E. P., Bui, C. K., Nett, J. E., Hartooni, N., Mui, M. C., Andes, D. R., Nobile, C. J. and Johnson, A. D. (2015). An expanded regulatory network temporally controls Candida albicans biofilm formation, *Molecular microbiology* 96(6): 1226–1239.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the american statistical association* 32(200): 675–701.

Gabaldón, T., Martin, T., Marcet-Houben, M., Durrens, P., Bolotin-Fukuhara, M., Lespinet, O., Arnaise, S., Boisnard, S., Aguileta, G., Atanasova, R. et al. (2013). Comparative genomics of emerging pathogens in the Candida glabrata clade, *BMC genomics* 14: 1–16.

Gabaldón, T., Naranjo-Ortíz, M. A. and Marcet-Houben, M. (2016). Evolutionary genomics of yeast pathogens in the Saccharomycotina, *FEMS yeast research* 16(6): fow064.

Gasch, A. P., Moses, A. M., Chiang, D. Y., Fraser, H. B., Berardini, M. and Eisen, M. B. (2004). Conservation and evolution of cis-regulatory systems in ascomycete fungi, *PLoS biology* 2(12): e398.

Glazier, V. E. (2022). EFG1, everyone's favorite gene in Candida albicans: A comprehensive literature review, *Frontiers in Cellular and Infection Microbiology* 12: 302.

Glazier, V. E., Kramara, J., Ollinger, T., Solis, N. V., Zarnowski, R., Wakade, R. S., Kim, M.-J., Weigel, G. J., Liang, S.-H., Bennett, R. J. et al. (2023). The Candida albicans reference strain SC5314 contains a rare, dominant allele of the transcription factor Rob1 that modulates biofilm formation and oral commensalism, *bioRxiv* .

Glazier, V. E., Murante, T., Murante, D., Koselny, K., Liu, Y., Kim, D., Koo, H. and Krysan, D. J. (2017). Genetic analysis of the Candida albicans biofilm transcription factor network using simple and complex haploinsufficiency, *PLoS genetics* 13(8): e1006948.

Gómez-Gaviria, M. and Mora-Montes, H. M. (2020). Current aspects in the biology, pathogeny, and treatment of Candida krusei, a neglected fungal pathogen, *Infection and drug resistance* pp. 1673–1689.

Hasan, F., Xess, I., Wang, X., Jain, N. and Fries, B. C. (2009). Biofilm formation in clinical Candida isolates and its association with virulence, *Microbes and infection* 11(8-9): 753–761.

Hsu, P.-C., Lu, T.-C., Hung, P.-H., Jhou, Y.-T., Amine, A. A., Liao, C.-W. and Leu, J.-Y. (2021). Plastic rewiring of Sef1 transcriptional networks and the potential of nonfunctional transcription factor binding in facilitating adaptive evolution, *Molecular Biology and Evolution* 38(11): 4732–4747.

Ihmels, J., Bergmann, S., Gerami-Nejad, M., Yanai, I., McClellan, M., Berman, J. and Barkai, N. (2005). Rewiring of the yeast transcriptional network through the evolution of motif usage, *Science* 309(5736): 938–940.

Jackson, A. P., Gamble, J. A., Yeomans, T., Moran, G. P., Saunders, D., Harris, D., Aslett, M., Barrell, J. F., Butler, G., Citiulo, F. et al. (2009). Comparative genomics of the fungal pathogens Candida dubliniensis and Candida albicans, *Genome research* 19(12): 2231–2244.

Jiang, S. and Mortazavi, A. (2018). Integrating ChIP-seq with other functional genomics data, *Briefings in functional genomics* 17(2): 104–115.

Johnson, A. D. (2017). The rewiring of transcription circuits in evolution, *Current opinion in genetics & development* 47: 121–127.

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes, *Genome research* 20(10): 1313–1326.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability, *Molecular biology and evolution* 30(4): 772–780.

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

Krassowski, T., Coughlan, A. Y., Shen, X.-X., Zhou, X., Kominek, J., Opulente, D. A., Riley, R., Grigoriev, I. V., Maheshwari, N., Shields, D. C. et al. (2018). Evolutionary instability of CUG-Leu in the genetic code of budding yeasts, *Nature communications* 9(1): 1887.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *The annals of mathematical statistics* 22(1): 79–86.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2, *Nature methods* 9(4): 357–359.

Lastauskienė, E., Čeputytė, J., Girkontaitė, I. and Zinkevičienė, A. (2015). Phenotypic switching of Candida guilliermondii is associated with pseudohyphae formation and antifungal resistance, *Mycopathologia* 179: 205–211.

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

Leng, P., Lee, P. R., Wu, H. and Brown, A. J. (2001). Efg1, a morphogenetic regulator in Candida albicans, is a sequence-specific DNA binding protein, *Journal of bacteriology* 183(13): 4090–4093.

Letunic, I. and Bork, P. (2021). Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation, *Nucleic acids research* 49(W1): W293–W296.

Lin, C.-H., Kabrawala, S., Fox, E. P., Nobile, C. J., Johnson, A. D. and Bennett, R. J. (2013). Genetic control of conventional and pheromone-stimulated biofilm formation in Candida albicans, *PLoS pathogens* 9(4): e1003305.

Liu, H., Styles, C. A. and Fink, G. R. (1996). Saccharomyces cerevisiae S288C has a mutation in FL08, a gene required for filamentous growth, *Genetics* 144(3): 967–978.

Lone, S. A. and Ahmad, A. (2019). Candida auris—the growing menace to global health, *Mycoses* 62(8): 620–637.

Luo, G., Wang, T., Zhang, J., Zhang, P. and Lu, Y. (2021). Candida albicans requires iron to sustain hyphal growth, *Biochemical and Biophysical Research Communications* 561: 106–112.

Ly, A., Verhagen, J. and Wagenmakers, E.-J. (2016). Harold jeffreys's default bayes factor hypothesis tests: Explanation, extension, and application in psychology, *Journal of Mathematical Psychology* 72: 19–32.

Mancera, E., Nocedal, I., Hammel, S., Gulati, M., Mitchell, K. F., Andes, D. R., Nobile, C. J., Butler, G. and Johnson, A. D. (2021). Evolution of the complex transcription network controlling biofilm formation in Candida species, *Elife* 10: e64682.

Mancera, E., Porman, A. M., Cuomo, C. A., Bennett, R. J. and Johnson, A. D. (2015). Finding a missing gene: EFG1 regulates morphogenesis in Candida tropicalis, *G3: Genes, Genomes, Genetics* 5(5): 849–856.

Marcos-Zambrano, L. J., Puig-Asensio, M., Pérez-García, F., Escribano, P., Sánchez-Carrillo, C., Zaragoza, O., Padilla, B., Cuenca-Estrella, M., Almirante, B., Martín-Gómez, M. T. et al. (2017). Candida guilliermondii complex is characterized by high antifungal resistance but low mortality in 22 cases of candidemia, *Antimicrobial agents and chemotherapy* 61(7): e00099–17.

Martchenko, M., Levitin, A. and Whiteway, M. (2007). Transcriptional activation domains of the Candida albicans Gcn4p and Gal4p homologs, *Eukaryotic cell* 6(2): 291–301.

Massey, S. E., Moura, G., Beltrão, P., Almeida, R., Garey, J. R., Tuite, M. F. and Santos, M. A. (2003). Comparative evolutionary genomics unveils the molecular mechanism of reassignment of the CTG codon in Candida spp., *Genome research* 13(4): 544–557.

Muñoz, J. F., Gade, L., Chow, N. A., Loparev, V. N., Juieng, P., Berkow, E. L., Farrer, R. A., Litvintseva, A. P. and Cuomo, C. A. (2018). Genomic insights into multidrug-resistance, mating and virulence in candida auris and related emerging species, *Nature communications* 9(1): 5346.

Nather, K. and Munro, C. A. (2008). Generating cell surface diversity in Candida albicans and other fungal pathogens, *FEMS microbiology letters* 285(2): 137–145.

Nielsen, F. (2019). On the Jensen–Shannon symmetrization of distances relying on abstract means, *Entropy* 21(5): 485.

Nobile, C. J., Andes, D. R., Nett, J. E., Smith Jr, F. J., Yue, F., Phan, Q.-T., Edwards Jr, J. E., Filler, S. G. and Mitchell, A. P. (2006). Critical role of Bcr1-dependent adhesins in C. albicans biofilm formation in vitro and in vivo, *PLoS pathogens* 2(7): e63.

Nobile, C. J., Fox, E. P., Nett, J. E., Sorrells, T. R., Mitrovich, Q. M., Hernday, A. D., Tuch, B. B., Andes, D. R. and Johnson, A. D. (2012). A recently evolved transcriptional network controls biofilm development in Candida albicans, *Cell* 148(1-2): 126–138.

Nobile, C. J. and Johnson, A. D. (2015). Candida albicans biofilms and human disease, *Annual review of microbiology* 69: 71–92.

Nocedal, I., Mancera, E. and Johnson, A. D. (2017). Gene regulatory network plasticity predates a switch in function of a conserved transcription regulator, *Elife* 6: e23250.

Nystrom, S. L. and McKay, D. J. (2021). Memes: A motif analysis environment in R using tools from the MEME Suite, *PLoS Computational Biology* 17(9): e1008991.

Pais, P., Costa, C., Cavalheiro, M., Romao, D. and Teixeira, M. C. (2016). Transcriptional control of drug resistance, virulence and immune system evasion in pathogenic fungi: a cross-species comparison, *Frontiers in Cellular and Infection Microbiology* 6: 131.

Panariello, B. H. D., Klein, M. I., Pavarina, A. C. and Duarte, S. (2017). Inactivation of genes TEC1 and EFG1 in Candida albicans influences extracellular matrix composition and biofilm morphology, *Journal of Oral Microbiology* 9(1): 1385372.

Patil, I. (2021). Visualizations with statistical details: The'ggstatsplot'approach, *Journal of Open Source Software* 6(61): 3167.

Payne, J. L. and Wagner, A. (2014). The robustness and evolvability of transcription factor binding sites, *Science* 343(6173): 875–877.

Pesole, G., Lotti, M., Alberghina, L. and Saccone, C. (1995). Evolutionary origin of nonuniversal cugser codon in some Candida species as inferred from a molecular phylogeny., *Genetics* 141(3): 903–907.

Price, M. N., Dehal, P. S. and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments, *PloS one* 5(3): e9490.

Priest, S. J. and Lorenz, M. C. (2015). Characterization of virulence-related phenotypes in Candida species of the CUG clade, *Eukaryotic cell* 14(9): 931–940.

Rajeh, A., Lv, J. and Lin, Z. (2018). Heterogeneous rates of genome rearrangement contributed to the disparity of species richness in Ascomycota, *BMC genomics* 19: 1–13.

Ramage, G., Vande Walle, K., Wickes, B. L. and López-Ribot, J. L. (2001). Biofilm formation by Candida dubliniensis, *Journal of Clinical Microbiology* 39(9): 3234–3240.

Ramos, L. S., Mello, T. P., Branquinha, M. H. and Santos, A. L. (2020). Biofilm formed by Candida haemulonii species complex: Structural analysis and extracellular matrix composition, *Journal of Fungi* 6(2): 46.

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

Satoh, K., Makimura, K., Hasumi, Y., Nishiyama, Y., Uchida, K. and Yamaguchi, H. (2009). Candida auris sp. nov., a novel ascomycetous yeast isolated from the external ear canal of an inpatient in a Japanese hospital, *Microbiology and immunology* 53(1): 41–44.

Seoighe, C. and Wolfe, K. H. (1998). Extent of genomic rearrangement after genome duplication in yeast, *Proceedings of the National Academy of Sciences* 95(8): 4447–4452.

Shen, X.-X., Steenwyk, J. L., LaBella, A. L., Opulente, D. A., Zhou, X., Kominek, J., Li, Y., Groenewald, M., Hittinger, C. T. and Rokas, A. (2020). Genome-scale phylogeny and contrasting modes of genome evolution in the fungal phylum Ascomycota, *Science advances* 6(45): eabd0079.

Sherry, L., Ramage, G., Kean, R., Borman, A., Johnson, E. M., Richardson, M. D. and Rautemaa-Richardson, R. (2017). Biofilm-forming capability of highly virulent, multidrug-resistant Candida auris, *Emerging infectious diseases* 23(2): 328.

Singh, R., Kaur, M., Chakrabarti, A., Shankarnarayan, S. A. and Rudramurthy, S. M. (2019). Biofilm formation by Candida auris isolated from colonising sites and candidemia cases, *Mycoses* 62(8): 706–709.

Skrzypek, M. S., Binkley, J., Binkley, G., Miyasato, S. R., Simison, M. and Sherlock, G. (2016). The candida genome database (CGD): incorporation of assembly 22, systematic identifiers and visualization of high throughput sequencing data, *Nucleic acids research* p. gkw924.

Teichmann, S. A. and Babu, M. M. (2004). Gene regulatory network growth by duplication, *Nature genetics* 36(5): 492–496.

Tuch, B. B., Li, H. and Johnson, A. D. (2008). Evolution of eukaryotic transcription circuits, *Science* 319(5871): 1797–1799.

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

Van Dongen, S. (2008). Graph clustering via a Discrete Uncoupling Process, *SIAM Journal on Matrix Analysis and Applications* 30(1): 121–141.

Voordeckers, K., Pougach, K. and Verstrepen, K. J. (2015). How do regulatory networks evolve and expand throughout evolution?, *Current opinion in biotechnology* 34: 180–188.

Wang, T. and Stormo, G. D. (2003). Combining phylogenetic data with co-regulated genes to identify regulatory motifs, *Bioinformatics* 19(18): 2369–2380.

Witchley, J. N., Penumetcha, P., Abon, N. V., Woolford, C. A., Mitchell, A. P. and Noble, S. M. (2019). Candida albicans morphogenesis programs control the balance between gut commensalism and invasive infection, *Cell host & microbe* 25(3): 432–443.

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

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L. et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data, *The Innovation* 2(3): 100141.

Wu, Y.-C., Rasmussen, M. D., Bansal, M. S. and Kellis, M. (2014). Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees, *Genome research* 24(3): 475–486.

Yue, H., Bing, J., Zheng, Q., Zhang, Y., Hu, T., Du, H., Wang, H. and Huang, G. (2018). Filamentation in Candida auris, an emerging fungal pathogen of humans: passage through the mammalian body induces a heritable phenotypic switch, *Emerging microbes & infections* 7(1): 1–13.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. et al. (2008). Model-based analysis of ChIP-Seq (MACS), *Genome biology* 9(9): 1–9.

Zuza-Alves, D. L., Silva-Rocha, W. P. and Chaves, G. M. (2017). An update on Candida tropicalis based on basic and clinical approaches, *Frontiers in microbiology* 8: 1927.

Chapter 3

Population genomic variants in *C. albicans* biofilm components

3.1 Abstract

Candida albicans is rich in genetic diversity driven by single nucleotide polymorphisms, small insertions and deletions and large structural variations. These different types of variants have different effects on the binding sites of biofilm regulators. To examine these effects, we identified variants in 190 *C. albicans* strains, spanning more than 14 clades and estimated their effects on the regulators and target genes in the biofilm regulatory network and their *cis*-regulatory regions. We also estimated the evolutionary forces acting on the components of the biofilm network. We found that transcription factor binding site losses are more prevalent than gains and that these are driven predominantly by single nucleotide polymorphisms and small indels. These binding site losses are not compensated by gains in other parts of the *cis*-regulatory regions. We also found that motifs of some biofilm regulators are gained more readily than others due to their mutational proximity to other, low-affinity binding sites. Transcription factors in general have higher rates of nonsynonymous polymorphisms compared to other genes in the *C. albicans* genome. These variants are found predominantly in intrinsically disordered regions, which are segments of DNA-binding proteins that play critical roles in localizing transcription factors to promoters. In addition to biofilm regulators, target genes upregulated in biofilm conditions are also variant-rich compared to genes upregulated in planktonic conditions. This suggests that intraspecific diversity in biofilm regulatory networks is driven by gene expression changes due to binding site losses and variants in highly expressed genes, which likely affects mRNA levels of these genes.

3.2 Introduction

Biofilm formation is highly varied not only between *Candida* species (Hawser and Douglas; 1994; Kuhn et al.; 2002; Villar-Vidal et al.; 2011; Pujol et al.; 2015), but also among strains (Li et al.; 2003; Jain et al.; 2007; Soll and Daniels; 2016; Huang et al.; 2019). Importantly, Huang et al. (2019) observed that the downstream effects of deleting biofilm regulators in *C. albicans* are dependent on their genetic backgrounds (Huang et al. (2019)). In addition to observed phenotypic differences of clinical and environmental strains of *C. albicans* (**?**), population structure (Ropars et al.; 2018) and genomic diversity driven by single nucleotide polymorphisms (SNPs) (Hirakawa et al.; 2015) and copy number variations (CNVs) (Bensasson et al.; 2019) have also been reported. Hirakawa et al. (2015) showed that genome-wide heterozygosity is correlated with increased fitness and hypothesized that adaptive genomic changes in clinical isolates are enriched for genes involved in the host response (Hirakawa et al. (2015)). Ropars et al. (2018) showed that a subspecies of *C. albicans*, *Candida africana*, has undergone extensive pseudogenization in genes involved in virulence and morphogenesis, possibly contributing to its restriction to a specific host environmental niche (Ropars et al. (2018)). A recent study that identified the genetic diversity of *C. albicans* strains within the same host observed high genetic diversity compared to *in vitro* strains, suggesting that the host environment likely poses unique adaptive pressures on *C. albicans* (Sitterlé et al. (2019)). Together, these studies suggest that genetic diversity is prevalent in isolates of *C. albicans* which likely affects biofilm regulatory network components, and their abilities to form biofilms, consequently leading to changes in virulence properties in host settings. In this chapter, we examine the source of genetic variants affecting biofilm network components and make predictions on the effects of these variants on biofilms.

3.2.1 Genetic diversity in *C. albicans*

C. albicans has a cryptic genomic life cycle, where it can exist as haploid, diploid or tetraploid at various stages. Some studies that focused on individual genes in the *C. albicans* genome observed that inheritance is largely clonal (Lott et al. (1999); Legrand et al. (2019)). Under some environmental conditions, however, diploid cells of different mating types can mate to form tetraploid zygotes, which undergo concerted genome loss to reach the diploid state (McManus and Coleman; 2014). This parasexual cycle generates high genetic diversity through recombination in the progeny of an otherwise clonal population (Berman and Hadany; 2012).This parasexual cycle is hypothesized to be an adaptive stress response, triggered to generate diverse phenotypes to cope with the stress, hence regulating evolvability (Berman and Hadany; 2012; Radman et al.; 1999). This process also enables *C. albicans* to tolerate large scale chromosomal changes including translocations, copy number variations (CNVs), and maintenance and loss of heterozygosity (Selmecki et al.; 2010; Todd et al.; 2019). Indels are also very common in heterozygous regions in *C. albicans* strains (Wang et al.; 2018), predominantly in intergenic regions (Jones et al.; 2004; Ene et al.; 2018). Large structural variants (SVs) are inherently generated during the *C. albicans* life cycle, hence, studying the effects of these variants on components of the biofilm regulatory networks will provide a broader perspective on the source of biofilm network variation. Additionally, the higher mutation frequencies observed in intergenic regions suggest increased variation in the transcription factor binding sites (TFBS). To study the extent of variation across TFBS, we examined the role of SNPs, indels and SVs on the *cis*-regulatory regions and target genes of the biofilm network.

Previous population genetic studies identified 23 clades of *C. albicans* (Ropars et al.; 2018; Gong et al.; 2023). In 2001, Tietz et al. (2001) proposed classifying clade 13 or the *C. africana* clade as a separate species, due to limitations in its abilities to undergo morphological transitions typical of *C. albicans* (Tietz et al.; 2001); however, phylogenetic comparison of ribosomal DNA sequences indicated that *C. africana* is indistinguishable from other *C. albicans* strains (Romeo and Criseo; 2011). Nonetheless, the taxonomic classification of *C. africana* remains highly debated. In addition to the morphological limitations of the *C. africana* clade, Ropars et al. (2018) also showed that this clade has undergone extensive genome reduction, resulting in the restriction of its strains to very specific host niches (Ropars et al.; 2018). In contrast to clade 13 (the *C. africana* clade), the phenotypic characteristics of strains belonging to clade 1 are broad. Clade 1 is the most commonly found *C. albicans* clade, and is geographically widely distributed (Odds et al.; 2007). It is also known to be the most commonly identified clade among commensal isolates (Odds et al.; 2007). Although, different *C. albicans* clades have been reported to have different propensities to cause infection (Odds et al.; 2007) and occupy different host niches (Ropars et al.; 2018), the ability to form biofilms does not appear to be significantly different between *C. albicans* clades (MacCallum et al.; 2009). We used a set of 190 strains from Hirakawa et al. (2015), Ropars et al. (2018) and Gong et al. (2023), belonging to 14 *C. albicans* clades for this study and identified SNPs, indels and SVs in each strain (Hirakawa et al.; 2015; Ropars et al.; 2018; Gong et al.; 2023).

3.3 Results

3.3.1 Diversity in *C. albicans* strains

We identified SNPs in 190 *C. albicans* strains using the Genome Analysis Toolkit (GATK) (McKenna et al.; 2010). We then inferred the phylogenetic relationships between strains using 572541 alignment sites as described in the methods section. Briefly, we used RAxML-NG (Kozlov et al.; 2019) with the General Time Reversible (GTR+G) model (Tavaré; 1986) with variable substitution rate at each site (Yang;

1994) as the substitution model. We used 20 starting trees for branch length estimation and 100 Felsenstein bootstraps (Felsenstein; 1985) to assess support for the branches. We visualized the phylogenetic tree (Figure 3.1) using iTOL (Letunic and Bork; 2021). The annotated the strains with clade designation from previous studies, Hirakawa et al. (2015); Ropars et al. (2018); Gong et al. (2023), through multilocus sequence typing (MLST) using seven housekeeping genes (Hirakawa et al.; 2015; Ropars et al.; 2018; Gong et al.; 2023). We identified 13 previously reported clades of *C. albicans* as well as *C. africana* (Figure 3.1) (Hirakawa et al.; 2015; Ropars et al.; 2018; Gong et al.; 2023). We also identified large SVs in these strains using GRIDSS (Cameron et al.; 2017, 2021).

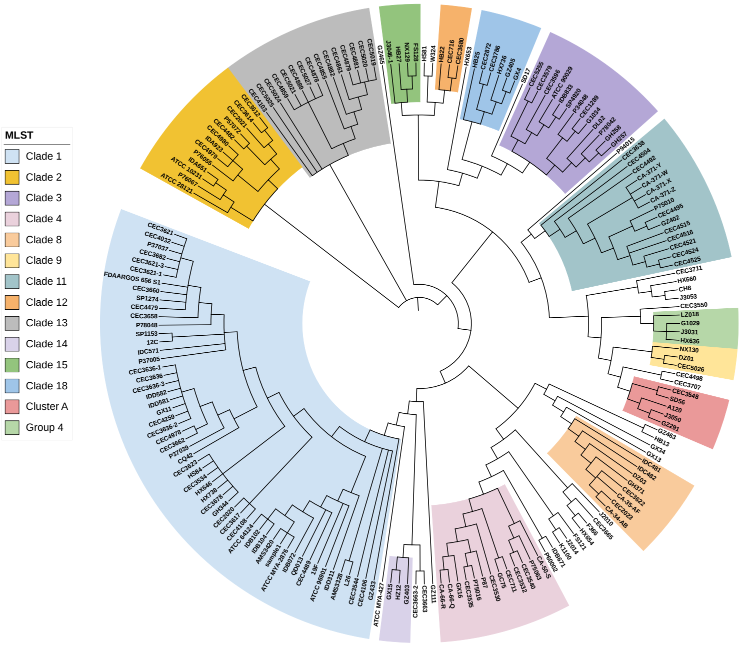
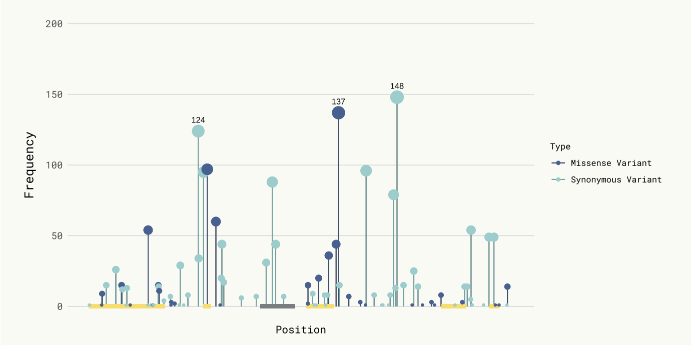
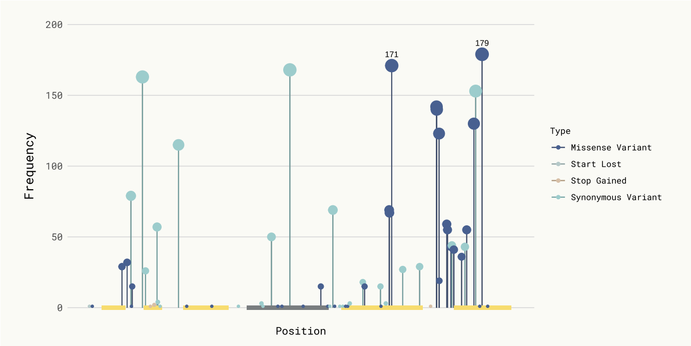


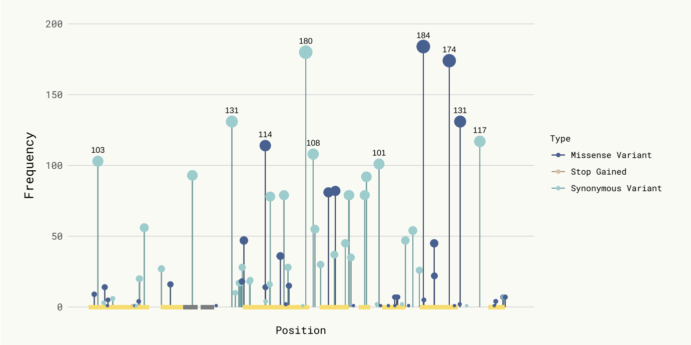
Figure 3.1: **Phylogenetic tree of the 190 strains used in this study.** Phylogenetic reconstruction of strains in this study is based on 572541 alignment sites using RAxML-NG (Kozlov et al.; 2019) with the GTR+G substitution model (Tavaré; 1986; Yang; 1994) using 20 starting trees and 100 bootstrap trees. The colors denote clades annotated by Hirakawa et al. (2015); Ropars et al. (2018); Gong et al. (2023) through multilocus sequence typing (MLST) using seven housekeeping genes. The tree is visualized using iTol (Letunic and Bork; 2021). Clade 1 also contains the ubiquitously used reference strain, *C. albicans* SC5314. Clade 13 is a subspecies of *C. albicans*, designated as *C. africana*. This clade is unique due to differences in morphology and host-niche occupancy of its strains compared to other *C. albicans* clades.



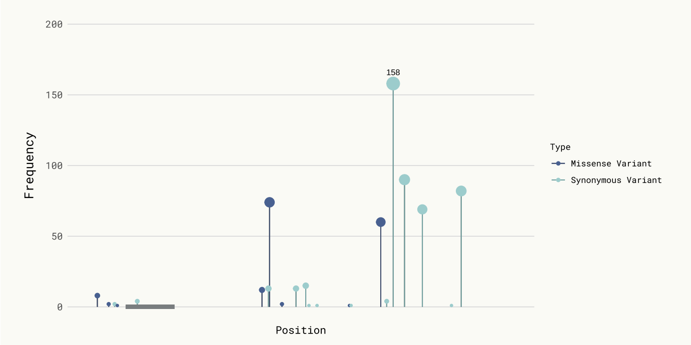
(a) *RFX2*



(b) *EFG1*



(c) *BCR1*



(d) *GAL4*

Figure 3.2: **High frequency non-synonymous variants are found in intrinsically disordered regions of biofilm regulators.**
Location of SNPs (x-axis) identified in the protein-coding gene. The y-axis is the frequency of strains harboring a SNP at the corresponding location. The colors represent the type of SNP. The yellow bars on the x-axis show intrinsically disordered regions (IDRs) and gray bars show the protein-protein interaction domain (PPI) or DNA-binding domain. All domain annotations were obtained from UniProt (The UniProt Consortium; 2023). IDRs are annotated in UniProt using MobiDB-lite (Necci et al.; 2020; Piovesan et al.; 2021) and PPIs and DNA-binding domains are annotated in UniProt using PROSITE (Sigrist et al.; 2012). *RFX2* harbors 92 unique SNPs, most of which are low-frequency mutations. Loss-of-function mutations are present in *EFG1* four strains. High frequency, non-synonymous mutations are abundant in the *IDR*s of *BCR1*. No IDRs are annotated for *GAL4* and this gene contains very few variants (22 unique variants).

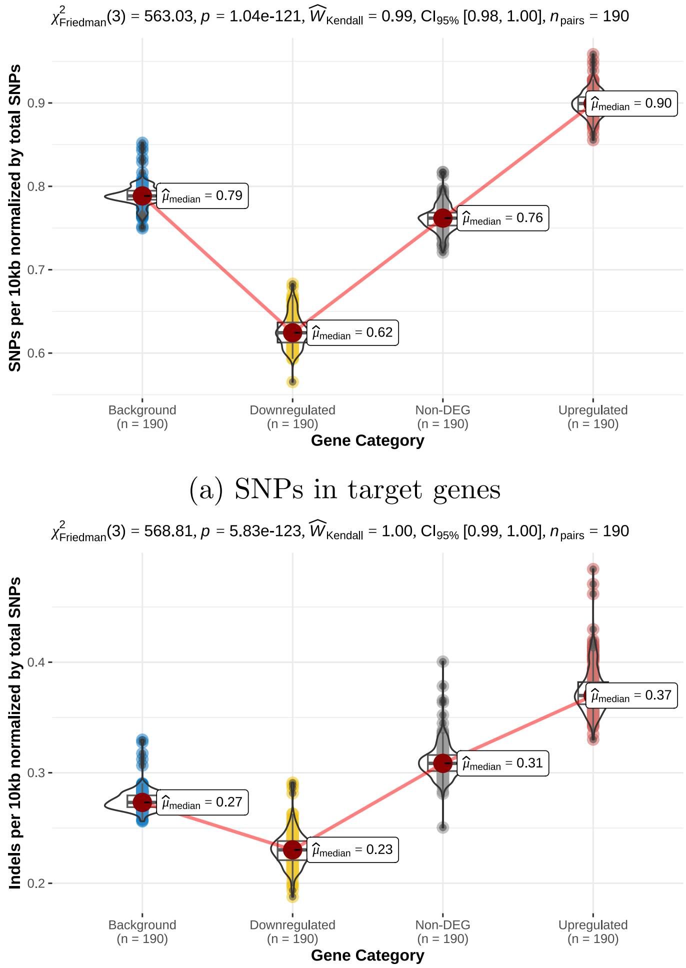
3.3.2 SNPs and indels are mostly found in intrinsically disordered regions of regulators

SNPs and indels in biofilm regulators occur at varying frequencies. The number of strains harboring mutations varied among biofilm regulators is shown in Figure 3.2a-Figure 3.2d. *RFX2* has the most number of identified SNPs (92 unique SNPs), but most of them are low-frequency mutations (Figure 3.2a). *BCR1* contains 84 unique SNPs (Figure 3.2c). Approximately 25% of these SNPs are present in at least 50 strains, indicating that these variants are prevalent in the population. We identified loss-of-function mutations in *EFG1*, specifically gain of a stop codon in four strains (Figure 3.2b). These variants truncate Efg1 upstream of the DNA-binding domain, rendering the protein nonfunctional. Among the strains containing this loss-of-function mutation is P94015, which has been reported to be biofilm defective (Hirakawa et al. (2015)). We identified very few mutations for *GAL4*. We found that biofilm regulators with one or more annotated IDRs harbor many SNPs and contain frameshift indels as well. These mutations, especially nonsynonymous mutations, are clustered in *IDR*s, suggesting that these regions are under weak purifying selection. Khan et al. (2015) report that inframe indels are also abundant in IDRs, at a rate similar to pseudogenes and non-coding regions. Afanasyeva et al. (2018) also identify rapid evolution of IDRs relative to structured regions, indicating that these regions accumulate mutations more frequently.

3.3.3 Genes upregulated during biofilm formation harbor more variants while their *cis*-regulatory regions are more conserved

We then tested whether the abundance of SNPs and indels was different between the target genes in the biofilm regulatory network and other genes in the genome. We classified the genes into the following groups: “Upregulated” if the gene is regulated by at least one biofilm regulator and upregulated ($\log_2$-fold change > 0.58 and FDR-adjusted p-value < 0.05) under biofilm conditions, “Downregulated” if the gene is regulated by at least one biofilm regulator and upregulated in planktonic conditions, and “Non-DEG” if the gene is regulated by at least one biofilm regulator and is not differentially expressed between biofilm and planktonic conditions. All other genes that are not regulated by any of the TFs are classified as “Background”. We compared the number of SNPs normalized by the total number of SNPs in each strain, for genes belonging to the categories mentioned above (Figure 3.3a). We compared indels in these gene categories as well (Figure 3.3b). We found that both SNPs and indels are less prevalent in genes upregulated in the planktonic cells. Whereas, SNPs and indels are more abundant in the genes upregulated during biofilm formation. We used the Friedman test (Friedman; 1937) to compare the difference in the variants between the four gene categories and observed significant differences in the abundance of both SNPs (Kendall’s W effect size=0.99) and indels (Kendall’s W effect size=1) between the categories. We also calculated the proportion of strains containing synonymous and nonsynonymous SNPs and the 95% confidence interval of mean proportion of strains containing these variants across gene categories (Table 3.1). We show that nonsynonymous SNPs are significantly less frequent in “Downregulated” genes.

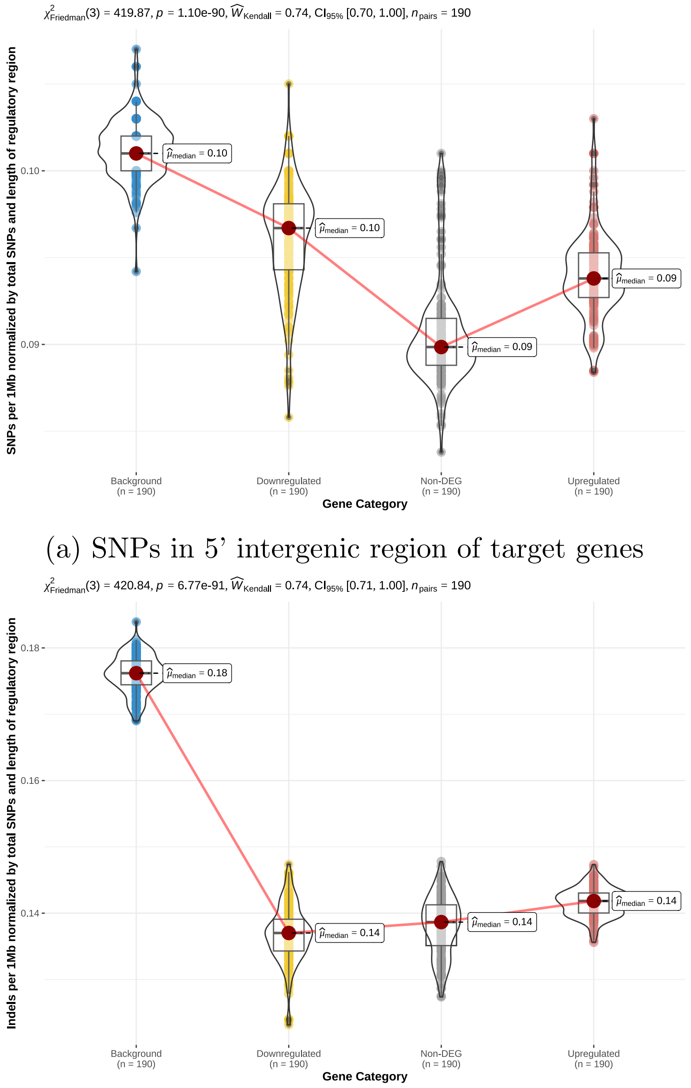
We then compared the abundance of SNPs and indels in the *cis*-regulatory region (5’ intergenic region) of the “Upregulated”, “Downregulated”, “Non-DEG” and “Background” genes (Figure 3.4). Interestingly, fewer SNPs and indels were found in the *cis*-regulatory region of genes upregulated under biofilm conditions. Conversely, variants were more abundant in the *cis*-regulatory regions of genes that are not in the biofilm regulatory network. Both indels (Kendall’s W effect size=0.74) and SNPs (Kendall’s W effect size=0.74) were less prevalent in the 5’ intergenic region of target genes in the biofilm regulatory network. This suggests that genes that are highly expressed in planktonic cells are likely under purifying selection. Genes expressed under biofilm conditions accumulate more mutations, while their corresponding *cis*-regulatory regions accumulate less.



(a) SNPs in target genes

(b) Indels in target genes

Figure 3.3: **Increased abundance of variants observed in genes upregulated under biofilm conditions.**
The number of SNPs and indels were normalized by the number of variants in each strain per 10 kb. Genes were categorized based on their expression during biofilm formation. Gene upregulated in biofilm formation ("Upregulated"), genes upregulated in planktonic cells ("Downregulated"), genes not differentially expressed between the two conditions ("Non-DEG") and genes that are not part of the biofilm regulatory network (indicated by category "Background") are shown. SNPs and indels are significantly more prevalent in genes highly expressed in biofilms. Conversely, genes highly expressed in planktonic cells contain less variants compared to genes that are "Background" and "Non-DEG".



(a) SNPs in 5' intergenic region of target genes

(b) Indels in 5' intergenic region of target genes

Figure 3.4: **Decreased abundance of variants were observed in *cis*-regulatory regions of genes upregulated under biofilm conditions.**
The number of SNPs and indels were normalized by the number of variants in each strain and by the cumulative length of intergenic regions in each gene category. Genes were categorized based on their expression during biofilm formation. Genes upregulated in biofilm formation ("Upregulated"), genes upregulated in planktonic cells ("Downregulated"), genes not differentially expressed between the two conditions ("Non-DEG") and genes that are not part of the biofilm regulatory network (indicated by category "Background") are shown. In contrast to variants in protein-coding genes, variants in *cis*-regulatory elements of genes in the biofilm *GRN* contain fewer SNPs and indels.

Gene Category	95% confidence interval of mean proportion of strains with synonymous variants	95% confidence interval of mean proportion of strains with nonsynonymous variants
Background	[0.791 - 0.808]	[0.716 - 0.737]
Downregulated	[0.760 - 0.804]	[0.592 - 0.654]
Non-DEG	[0.777 - 0.816]	[0.678 - 0.727]
Upregulated	[0.805 - 0.833]	[0.730 - 0.763]

Table 3.1: Proportion of strains with synonymous and nonsynonymous variants in gene categories.

3.3.4 Variants in *cis*-regulatory region drive TFBS gains for some TFs

Given that we observed fewer variants in the *cis*-regulatory regions of target genes in the biofilm network, we wanted to assess if the observed SNPs and indels result in gains or losses of putative TFBSs in these regions. We searched for the presence of TF binding motifs obtained in chapter 2 for each strain, and computed the loss-to-gain ratio for each motif in the strain. Previously, we identified a primary and alternate motif for each TF in the *C. albicans* biofilm network. The Brg1 alternate motif was similar to another TF in the regulatory network, Ndt80. Hence, this motif was removed. We then used FIMO (Grant et al.; 2011), a motif search tool to identify motif presence in the intergenic regions of the strains. FIMO requires a motif of at least seven base pairs in length, hence the primary motif of Brg1 (6 bps) and alternate motif of Rob1 (6 bps) were removed from further analysis.

In general, TFBSs are more frequently lost than gained, but the loss-to-gain ratio varies by the motif. The distribution of loss-to-gain ratio for each TF across strains is shown in Figure 3.5a. We used the Kruskal-Wallis nonparametric test (Kruskal and Wallis; 1952) to estimate whether the median loss-to-gain ratio varies between TFs. We found that the ordinal effect size, measured by ϵ^2, was 0.17 indicating a moderate influence of the TF on variation in the TFBS loss-to-gain ratio. We also found that the loss-to-gain ratio of Efg1 and Tec1 binding motifs have a bimodal distribution and wanted to verify if this ratio was different between the primary and alternate motifs. For Efg1, the primary and alternate motif have very distinct loss-to-gain ratios (Figure 3.5b). The primary motif has a lower loss-to-gain ratio compared to the alternate motif. Since alternate motifs have different loss-to-gain distribution, we restricted out comparison to the primary TF motifs (Figure 3.5c). We found that the ordinal effect size, ϵ^2, was 0.49 indicating a strong influence of the TF primary motifs on variation in the TFBS loss-to-gain ratio. Despite excluding the alternate motif, the loss-to-gain ratio for the Tec1 motif still followed a bimodal distribution.

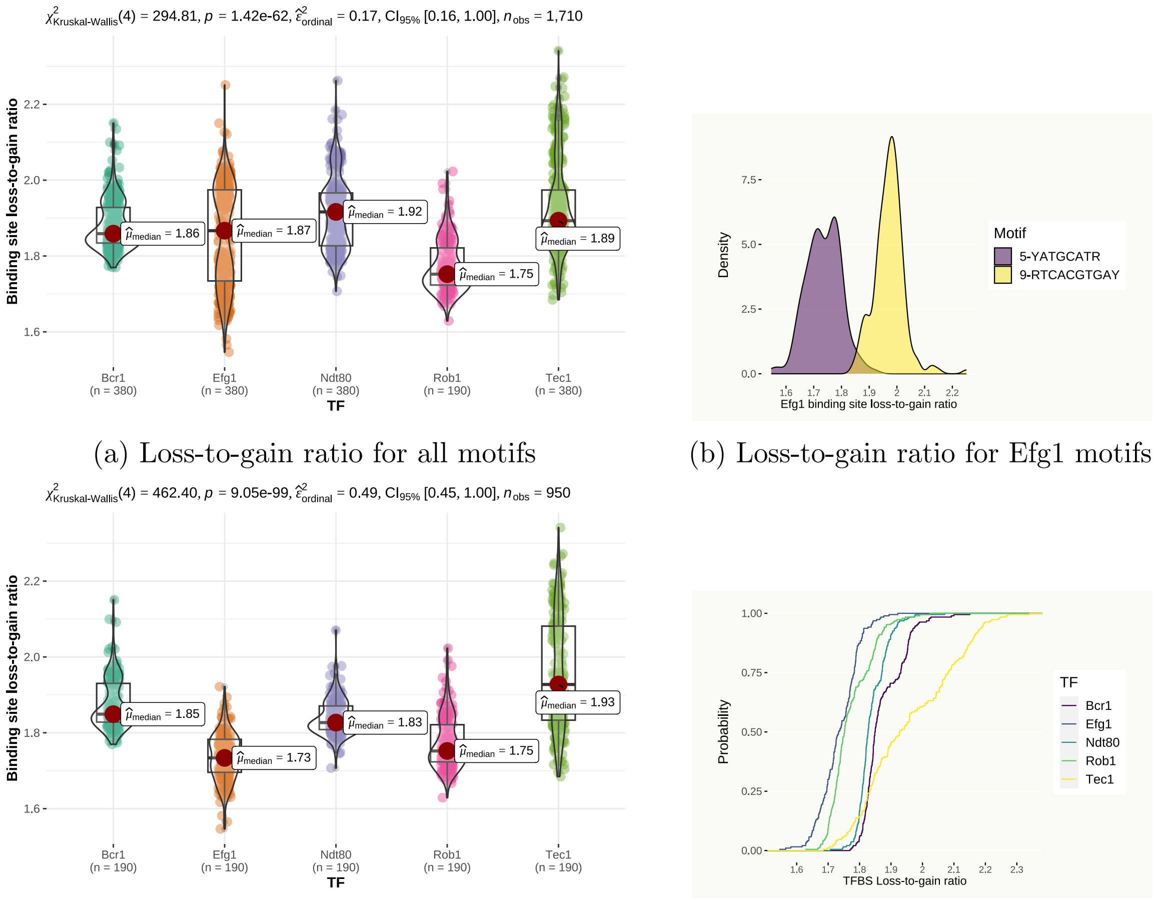
This shows that intraspecific variation is higher in Tec1 motifs. We examined the strains with increased loss-to-gain ratios of Tec1 motifs and found that these strains belong to clade 1. Since, the variance of loss-to-gain ratios is heterogeneous between TFs, we performed the Anderson-Darling test (Scholz and Stephens; 1987) to verify whether the distributions are similar (Figure 3.5d). We found significant differences (Anderson-Darling statistic=180.1) in loss-to-gain ratios between the primary motifs of the TFs. The loss-to-gain ratios of Rob1 and Efg1 motifs are lower than other TF motifs, driven by TFBS gains for these two motifs. Conversely, the loss-to-gain ratio of Tec1 is higher than other TF motifs.

3.3.5 TFBS losses and gains are balanced in structural variants

We observed that SNPs and small indels (less than 10 bp) resulted in losses of TFBS for all TFs. We wanted to determine whether SVs have a similar impact on TFBSs. We identified SVs as described in chapter 4 and used custom methods to filter and annotate the SVs. We then identified instances of TFBS losses in deletions, and TFBS gains in insertions and duplications. We hypothesized that TFBS gains or losses due to inversions will occur only at the break-ends of inversion events, so we did not estimate the effect of inversions on the TFBSs. The frequencies of insertions were comparable to deletions across strains but the deletions were typically longer (Figure 3.6b). Large duplications and inversions were also identified in the strains (Figure 3.6b). We used the Kruskal-Wallis nonparametric test to estimate whether the median lengths of different types of SV events vary and found that the difference in median lengths is moderately explained by the SV type (ordinal effect size ϵ^2=0.1). We then estimated the number of TFBS gains using FIMO (Grant et al.; 2011) to detect possible binding occurrences for the primary and alternate TF motifs as described in the previous section. We found that TFBS losses and gains are balanced in the SVs (Figure 3.6a) and the loss-to-gain ratios are approximately one for all TFs. The TFBS losses driven by large deletions are compensated by large duplications. Duplications occurred at lower frequencies in the *C. albicans* strains, but the duplication sizes are large (median length is 630 bp). Insertion lengths are quite small (median length is 12 bp) to compensate completely for the TFBS losses. This shows that the primary mode of TFBS loss is driven by SNPs and smaller indels.

3.3.6 Variants driving reduction in motif strength of Efg1 and Rob1 are prevalent in the population

We compared binding site loss, gain and conservation driven by SNPs and indels in 190 strains, against the common *C. albicans* lab reference strain SC5314. We



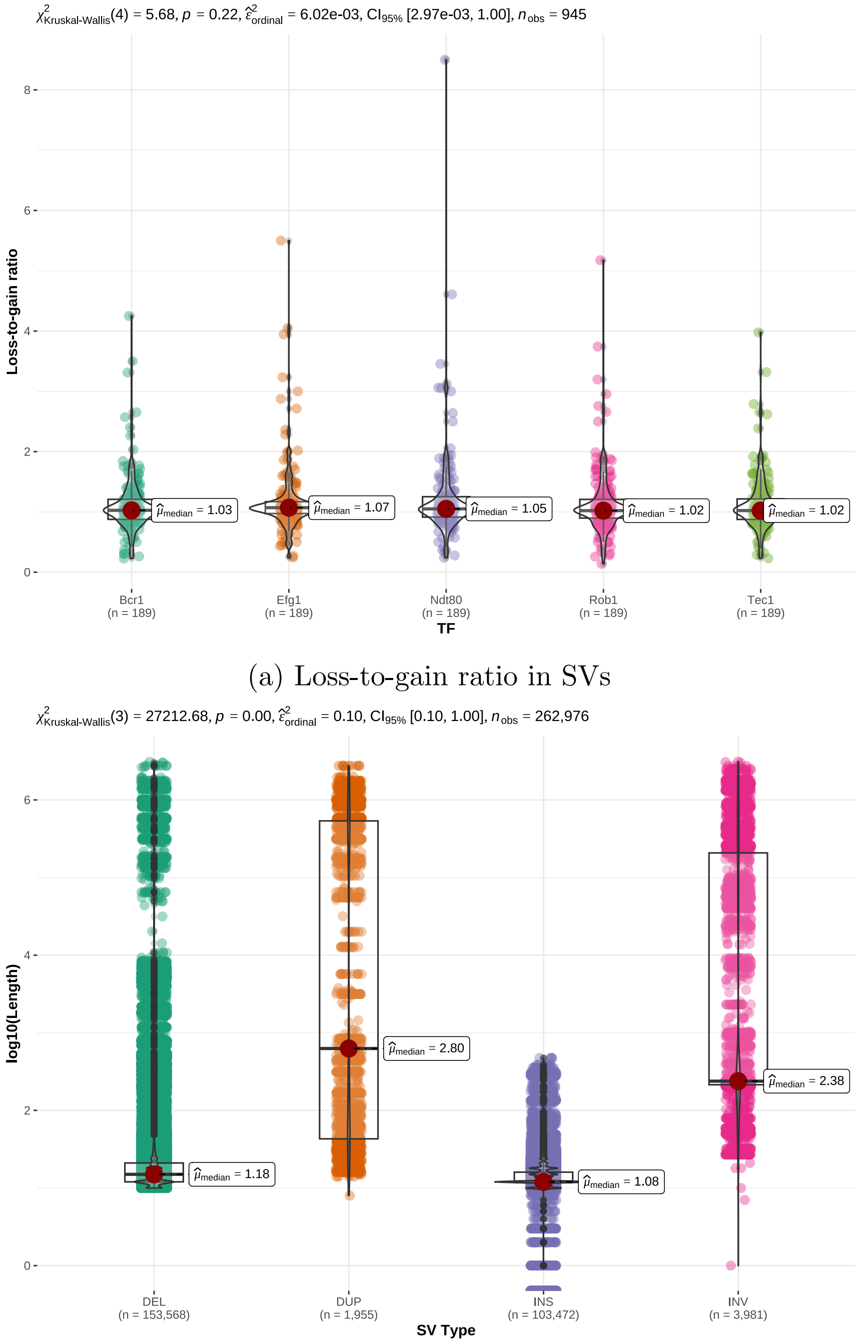
(a) Loss-to-gain ratio for all motifs

(b) Loss-to-gain ratio for Efg1 motifs

(c) Loss-to-gain ratio for primary motifs

(d) Loss-to-gain ratio (CDF)

Figure 3.5: **Intraspecific variants drive binding site loss for TFs, but binding site gains are observed for some TFs.**
The genome-wide loss-to-gain ratio was estimated for each binding motif. (a) The loss-to-gain ratio varies by the motif. The distribution of loss-to-gain ratio for each TF across strains is shown here. ϵ^2=0.17 indicating a moderate influence of the TF on variation in TFBS loss-to-gain ratio. Loss-to-gain ratios of Efg1 and Tec1 binding motifs have a bimodal distribution. (b) The primary and alternate Efg1 motifs have very distinct loss-to-gain ratios. (c) The ordinal effect size, ϵ^2=0.49, indicating a strong influence of the TF primary motifs on variation in TFBS loss-to-gain ratios. (d) The CDF of loss-to-gain ratios the primary motifs. The loss-to-gain ratios of Rob1 and Efg1 motifs are lower and the loss-to-gain ratio of Tec1 is higher compared to other TF motifs.



(a) Loss-to-gain ratio in SVs

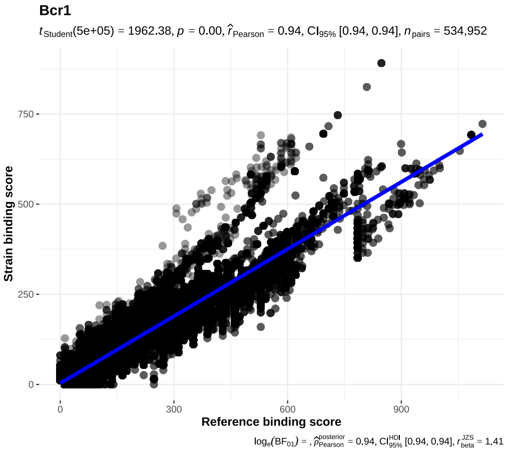
(b) SV length distribution across strains

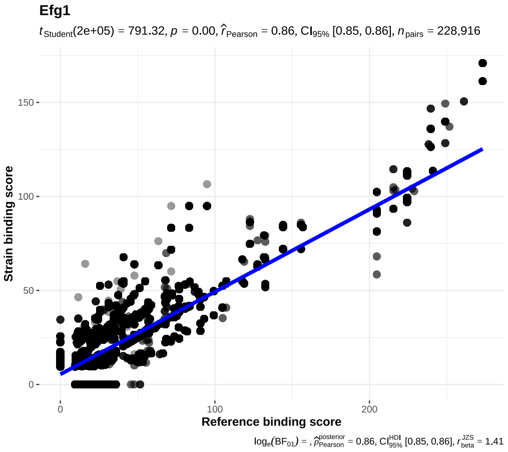
Figure 3.6: **TFBS loss-to-gain ratios in SVs are close to one for all TFs.** SVs were identified using GRIDSS (Cameron et al.; 2017, 2021). Binding site loss due to deletions and binding site gain in insertions and duplications was estimated for each TF. (a) The loss-to-gain ratios are approximately one for all TFs. The x-axis denotes the TF and y-axis the loss-to-gain ratios in deletions, duplications and insertions. (b) TFBS gain is driven primarily by duplications compared to insertions. The x-axis denotes the type of SV and y-axis the log_{10}(length) of the SV. SV data is aggregated across all strains.

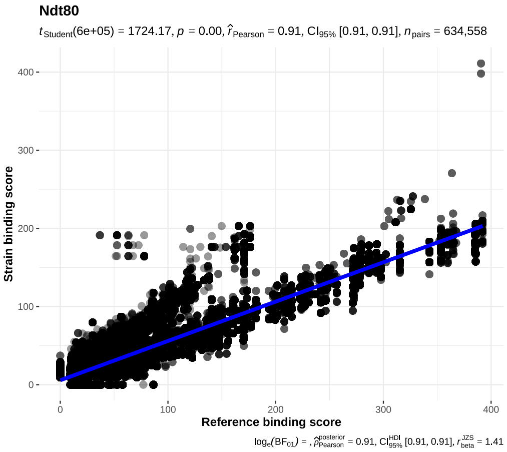
found that TFBS loss was the most prevalent with a median binding site loss of 1604 across strains. Conservation/retention of binding sites were the second commonly occurring events following TFBS loss, with a median of 1237 retained sites across strain. TFBS gains occur less frequently, with a median gain of 854.5. Previously, we showed that loss-to-gain ratios varied between the TFs. We then compared the motif strength in the predicted binding site to understand if motif strength is conserved across TFs (Figure 3.6). We computed the Pearson correlation between the predicted motif strength in SC5314 and the strains for all conserved TFBSs. We found that motif strength is highly correlated for Bcr1 (Pearson's r = 0.94), Ndt80 (Pearson's r = 0.91) and Tec1 (Pearson's r = 0.95). The Efg1 and Rob1 motif strengths were less correlated between SC5314 and the strains; the Pearson's r was 0.86 and 0.78, respectively. The slopes of the regression lines for Efg1 and Rob1 were 0.439 and 0.429, respectively, indicating lower motif strength in the strains compared to the reference SC5314. This shows that even when the number of binding sites are conserved, the motif strength is lower for Efg1 and Rob1 binding sites in the strains compared to other TFs. This coupled with our previous finding that TFBS gains are observed more frequently for Efg1 and Rob1 compared to other TFs indicates that the mutational accessibility of lower affinity binding sites drives motif gains for these TFs. Another possibility is that these regulatory regions are under positive selection, resulting in prevalence of variants that result in binding site gains for the TFs.

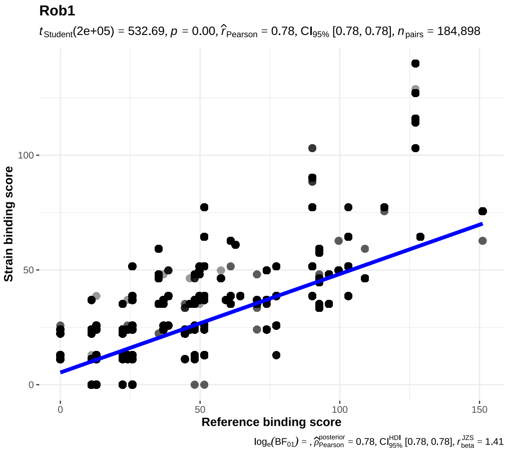
3.3.7 Selection pressure in components of the biofilm regulatory network

We examined the role of selection on *cis*-regulatory elements, and protein-coding genes. Specifically, we used Tajima's D (Tajima; 1989) to verify if the *cis*-regulatory regions of target genes in the biofilm regulatory network were under positive or purifying selection. We did not find a significant difference between the *cis*-regulatory network of target genes in the network compared to other genes in the *C. albicans* genome. We also did not find a difference in Tajima's D for *cis*-regulatory regions where multiple TFs bind, compared to the *cis*-regulatory regions regulated by a single TF. This was surprising, since we were expecting *cis*-regulatory regions with binding


Bcr1
$t_{\text{Student}}(5e+05) = 1962.38$, $p = 0.00$, $\hat{r}_{\text{Pearson}} = 0.94$, $\text{CI}_{95\%}$ [0.94, 0.94], $n_{\text{pairs}} = 534{,}952$
Strain binding score
0
250
500
750
0
300
600
900
Reference binding score
$\log_e(\text{BF}_{01}) = $, $\hat{\rho}^{\text{posterior}}_{\text{Pearson}} = 0.94$, $\text{CI}^{\text{HDI}}_{95\%}$ [0.94, 0.94], $r^{\text{JZS}}_{\text{beta}} = 1.41$


Efg1
$t_{\text{Student}}(2e+05) = 791.32$, $p = 0.00$, $\hat{r}_{\text{Pearson}} = 0.86$, $\text{CI}_{95\%}$ [0.85, 0.86], $n_{\text{pairs}} = 228{,}916$
Strain binding score
0
50
100
150
0
100
200
Reference binding score
$\log_e(\text{BF}_{01}) = $, $\hat{\rho}^{\text{posterior}}_{\text{Pearson}} = 0.86$, $\text{CI}^{\text{HDI}}_{95\%}$ [0.85, 0.86], $r^{\text{JZS}}_{\text{beta}} = 1.41$


Ndt80
$t_{\text{Student}}(6e+05) = 1724.17$, $p = 0.00$, $\hat{r}_{\text{Pearson}} = 0.91$, $\text{CI}_{95\%}$ [0.91, 0.91], $n_{\text{pairs}} = 634{,}558$
Strain binding score
0
100
200
300
400
0
100
200
300
400
Reference binding score
$\log_e(\text{BF}_{01}) = $, $\hat{\rho}^{\text{posterior}}_{\text{Pearson}} = 0.91$, $\text{CI}^{\text{HDI}}_{95\%}$ [0.91, 0.91], $r^{\text{JZS}}_{\text{beta}} = 1.41$


Rob1
$t_{\text{Student}}(2e+05) = 532.69$, $p = 0.00$, $\hat{r}_{\text{Pearson}} = 0.78$, $\text{CI}_{95\%}$ [0.78, 0.78], $n_{\text{pairs}} = 184{,}898$
Strain binding score
0
50
100
0
50
100
150
Reference binding score
$\log_e(\text{BF}_{01}) = $, $\hat{\rho}^{\text{posterior}}_{\text{Pearson}} = 0.78$, $\text{CI}^{\text{HDI}}_{95\%}$ [0.78, 0.78], $r^{\text{JZS}}_{\text{beta}} = 1.41$

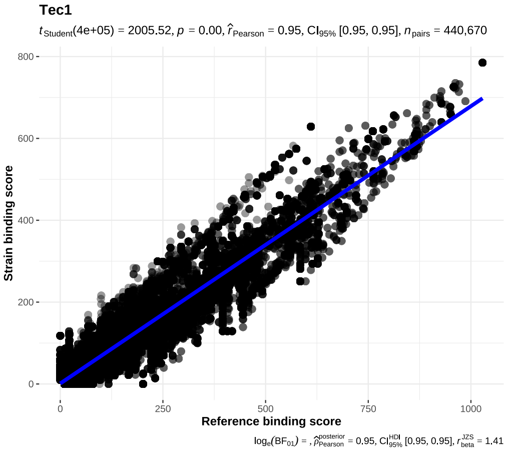

Tec1
$t_{\text{Student}}(4e+05) = 2005.52$, $p = 0.00$, $\hat{r}_{\text{Pearson}} = 0.95$, $\text{CI}_{95\%}$ [0.95, 0.95], $n_{\text{pairs}} = 440{,}670$
Strain binding score
0
200
400
600
800
0
250
500
750
1000
Reference binding score
$\log_e(\text{BF}_{01}) = $, $\hat{\rho}^{\text{posterior}}_{\text{Pearson}} = 0.95$, $\text{CI}^{\text{HDI}}_{95\%}$ [0.95, 0.95], $r^{\text{JZS}}_{\text{beta}} = 1.41$

Figure 3.6: **Binding strengths of TFs are correlated at conserved binding sites.**
For binding sites that are conserved in both the reference SC5314 and the *C. albicans* strains, the strength of the binding site is denoted by the sequence similarity of the binding site to the consensus motif identified using the reference strain. The strength of the binding site in the reference is shown in the x-axis and the strength of the corresponding binding site in the strain is shown in the y-axis. Motif strengths are highly correlated for Bcr1, Ndt80 and Tec1. Efg1 and Rob1 motif strength is weaker and less correlated between the reference strain SC5314 and other *C. albicans* strains, compared to other TFs.

sites for multiple TFs to be indicative of cooperative binding of TFs, which would be more evolutionarily constrained. One possible explanation for this unexpected finding is potential bias in our Tajima's D estimates. Tajima's D estimates are known to be biased when the recombination rate is high (Booker et al.; 2020) or when population bottlenecks are frequent (Gattepaille et al.; 2013; Fijarczyk and Babik; 2015). *C. albicans* undergoes population bottlenecks (Ene et al.; 2021) and its genome is characterized by high rates of recombination (Wang et al.; 2018; Anderson et al.; 2019). Hence, accounting for these events may uncover differences in selection pressures in the *cis*-regulatory regions (Booker et al.; 2020).

We also calculated the rate of nonsynonymous polymorphisms (dN) and rate of synonymous polymorphisms (dS) to compare the ratio (dN/dS) for protein-coding genes. dN/dS is conventionally used to compare rates of nonsynonymous substitutions against synonymous substitutions between species. For interspecific comparisons using this ratio, $dN/dS = 1$ implies that the gene is under neutral selection, $dN/dS > 1$ is indicative of positive selection and $dN/dS < 1$ is indicative of negative selection. When measuring dN/dS within a population, this ratio could be overestimated (Kryazhimskiy and Plotkin; 2008). Therefore, instead of using the absolute dN/dS to compare genes, we performed a gene set enrichment analysis (GSEA) by ranking the genes by their dN/dS ratio (Figure 3.7). We found that transcription factors have higher dN/dS ratios compared to other genes. We also found that genes involved in biofilm-related functional categories such as “interspecies interaction” and “filamentous growth” have elevated dN/dS ratios. The enrichment of these categories is also driven primarily by TFs, indicating TFs have a higher rate of nonsynonymous polymorphisms. Overall, only 43 genes had dN/dS values greater than 1. This indicates that most of the *C. albicans* genes in the genome are evolving under purifying selection, while genes involved in biofilm formation are also evolving under purifying selection, but have greater dN/dS values compared to the rest of the genes in the genome.

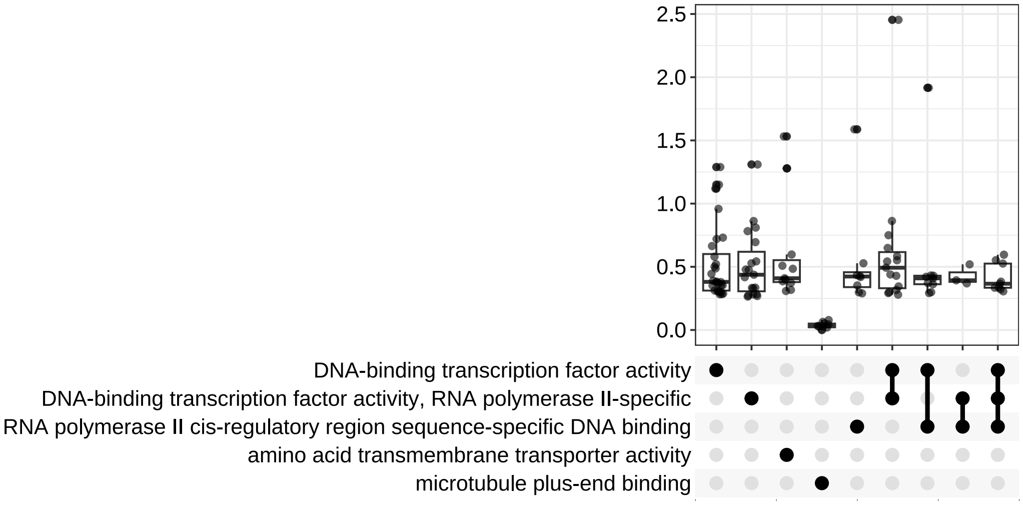


Figure 3.7: **TFs have elevated dN/dS ratios compared to other genes in the *C. albicans* genome.**
GSEA was performed by ranking protein-coding genes by their dN/dS ratios. The molecular functions comprising genes with elevated dN/dS ratios includes the functional category "DNA-binding transcription factor activity".

3.4 Discussion

In this study we found that variants are abundant in IDRs. Other studies have made similar observations in other yeast species such as *Saccharomyces cerevisiae* and *S. paradoxus* (Nilsson et al.; 2011; Khan et al.; 2015), *Drosophila* (Ridout et al.; 2010) and mammals (Khan et al.; 2015; Afanasyeva et al.; 2018). Specifically, Nilsson et al. (2011) show that IDRs are evolving under positive selection compared to other structural domains of proteins (Nilsson et al.; 2011). IDRs are unstructured regions of proteins, that are devoid of hydrophobic residues and rich in charged amino acids (Oldfield and Dunker; 2014; Mao et al.; 2010). These regions play a vital role in localizing TFs to promoters (Brodsky et al.; 2020). Importantly, Brodsky et al. (2020) show that the entirety of the IDR is required to target TFs to promoters and truncation of these regions result in partial loss of targets. IDR-rich TFs in *C. albicans* form phase-separated condensates (Frazer et al.; 2020). In higher eukaryotes, such condensates are known to recruit transcriptional machinery. Hence, the high prevalence of variants in IDRs suggest flexibility in interaction between TFs. One caveat to note is that IDRs can tolerate mutations while maintaining functional interactions with other proteins (Hultqvist et al.; 2017). Hence, they might still be evolving more neutrally compared to other domains in the protein, which might be evolving under purifying selection.

We found that target genes upregulated under biofilm conditions are variant-rich compared to the genes upregulated under planktonic conditions. This is likely due to either weaker purifying selection or balancing selection in the genes upregulated during biofilm formation. A recent study showed that genes expressed in later stages of the life history of invertebrates and vertebrates experience weak purifying selection, resulting in increased frequencies of observed deleterious mutations (Cheng and Kirkpatrick; 2021). If we equate this to biofilm formation as a developmental process, we can expect that genes highly expressed in mature biofilms could be evolving under weaker purifying selection compared to genes highly expressed in planktonic cells. Balancing selection is an alternate possible explanation for this observation. Even deleterious mutations observed in *EFG1*, for example, result in increased fitness of *C. albicans* as a commensal (Hirakawa et al.; 2015). Additionally, when balancing selection is in play, heterozygotes have higher fitness, compared to homozygotes in the population. In *C. albicans*, higher levels of heterozygosity is associated with increased fitness (Hirakawa et al.; 2015).

We observed increased TFBS loss among the *C. albicans* strains (53.7%-73%) compared to TFBS gains (23.1%-46.3%). Previously, some studies have estimated that binding site turnover is common between species (Dermitzakis and Clark; 2002; Ludwig et al.; 2000; Moses et al.; 2006), where binding site losses are compensated by gains elsewhere in the promoter region. We observed, however, an increase in binding site loss without a corresponding compensatory gain. A recent study comparing TFBS gains and losses between two yeast species, *S. cerevisiae* and *S. paradoxus*, found that losses and gains are more frequent than binding site turnover (Krieger et al.; 2022). In our intraspecific comparison, we found that TFBS losses are the most common, followed by turnover (where the losses are compensated by gains), and then conservation of the *cis*-regulatory region. TFBS gains occur at lower frequency and are dependent on the TF motifs. Binding site gains for Efg1 and Rob1 are higher than other TFs. We compared the mutational landscape of these binding motifs (Figure 3.8) with the landscape of the Ndt80 motif, which was gained at lower frequencies among strains. We found that the Efg1 (Figure 3.8a) and Rob1 (Figure 3.8b) landscapes are more rugged with multiple accessible, lower affinity binding sites, whereas the Ndt80 (Figure 3.8c) landscape is more smooth with a single, broad peak. The mutational landscape is consistent with our observation of increased gains in Efg1 and Rob1 as well as decreased correlation between motif strengths. The narrow peak, especially for Efg1 suggests that mutations in the binding site are more likely to result in lower motif strength, compared to Ndt80. Since Ndt80 has a broader peak, mutational neighbors are likely to have similar motif strength, hence the higher correlation between motif strength in the reference SC5314 and other *C. albicans* strains.

It is more likely that the TFBS gains and losses are driven by neutral selection. The determining factor of the consequence of the mutation is dependent on the binding motif and the mutational landscape of the motif, where some motifs are more tolerant of variations compared to others. We estimate the effect of variants on binding using the position probability matrices (PPMs) of the motifs. One disadvantage of using PPMs is that they assumes that positions in the motifs are independent of each other and do not account for epistasis and positional dependency of nucleotides. Although recent studies show that PPMs are predictive of binding affinity to alternative alleles (Boytsov et al.; 2022), other factors such as DNA shape (Schnepf et al.; 2020) and motif context (Avsec et al.; 2021) can affect binding. Hence, the TFBS losses that we observe are likely underestimates. Additionally, we note that not all binding sites are functional and we would need additional information such as gene expression and/or *in vivo* binding data from ChIP-Seq experiments to accurately predict the effect of the *cis*-regulatory variants.

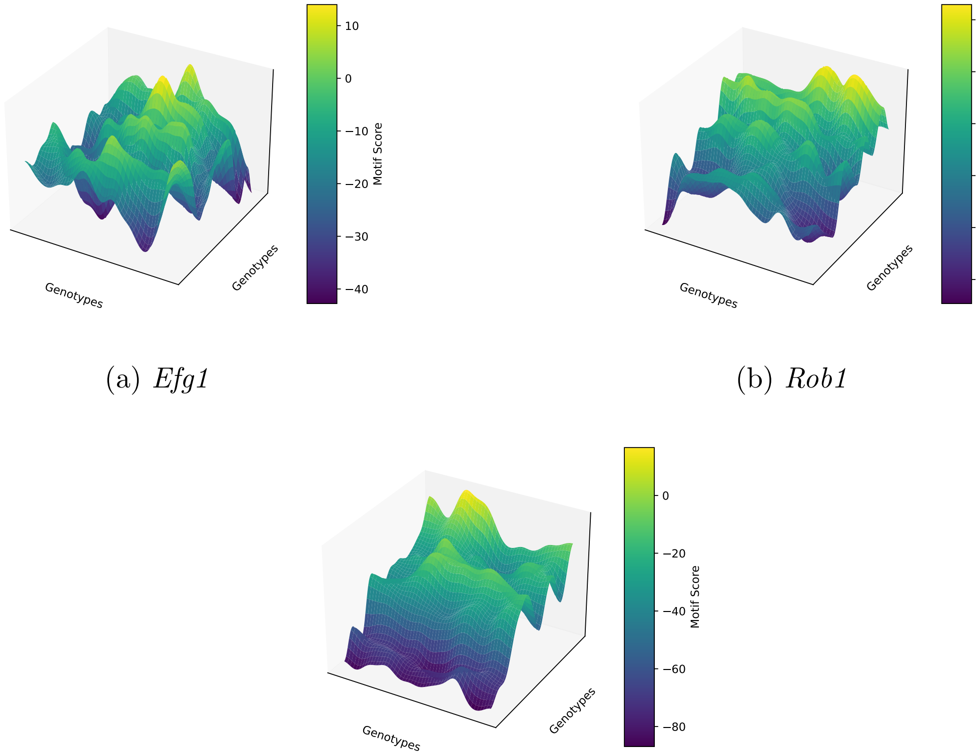
There is one other notable limitation in our estimates of dN/dS and Tajima's D, measuring the effect of selection on protein-coding and non-coding regions, respectively. As mentioned earlier, dN/dS is conventionally used for estimating nonsynonymous and synonymous substitutions between species. Studies, however, have used dN/dS to measure intraspecific nonsynonymous and synonymous polymorphisms in bacteria (Fleischmann et al.; 2002; Feil et al.; 2003) and laboratory evolved *S. cerevisiae* strains (Johnson et al.; 2021). When we use population genetic approaches to estimate dN/dS, it is more likely to be biased to assumptions on population size and demography (Kryazhimskiy and Plotkin; 2008). Additionally, variation in sequencing depth can also introduce errors when comparing these statistics between genomic regions. As part of future work, we plan to address these biases by accounting for read depth variation between different sites and strains (Korneliussen et al.; 2013).

3.5 Methods

3.5.1 Data acquisition

To identify variants segregating in the *C. albicans* population, we obtained 190 publicly available, short-read, genome sequences for *C. albicans* clinical strains from previous studies (Hirakawa et al.; 2015; Ropars et al.; 2018; Li et al.; 2022; Gong et al.; 2023). These sequences were obtained for samples from individuals with superficial and systemic *Candida* infections, as well as commensal strains from healthy individuals. We used these 190 geographically distributed strains to estimate genetic diversity in components of the biofilm GRN. Raw, paired-end reads were retrieved using the sra-toolkit (version 3.0.3) from National Center for Biotechnology Information (NCBI).

GRN components were comprised of master regulators of biofilm formation (Bcr1, Brg1, Efg1, Ndt80, Rob1 and Tec1), the target genes they bind to and the upstream (5') intergenic region of the target genes. An updated biofilm regulatory network was constructed using ChIP-seq data from Mancera et al. (2021) as described in chapter 2.



(a) *Efg1*

(b) *Rob1*

(c) *Ndt80*

Figure 3.8: **Mutational landscapes of Efg1, Rob1 and Ndt80 primary motifs.** The mutational landscapes of Efg1 and Rob1 are more rugged with many accessible peaks compared to Ndt80. The x-axis and y-axis indicate all possible k-mers, where k is the motif length and the z-axis denotes the motif score calculated from the PPM of the motif. The breadth of the global peak corresponds to mutational robustness.

Briefly, we obtained single-end raw reads from ChIP-Seq experiments published by Mancera et al. (2021) from the NCBI Gene Expression Omnibus (GEO). We aligned the reads to the *C. albicans* reference genome SC5314 using Bowtie2 (version 2.2.5) (Langmead and Salzberg (2012)). A custom peak-calling workflow was developed to increase sensitivity in the peak calling step using MACS2 (version 2.2.7.1) (Zhang et al. (2008)). A custom python script was used to annotate the peaks and retrieve sequences to identify binding motifs using STREME from the MEME-suite (version 5.5.2) (Bailey (2021)) and the memes package (version 1.2.5) in R (Nystrom and McKay (2021)). The workflow is described in detail in chapter 4. Gene expression of target genes in the updated GRN was obtained from Nobile et al. (2012). Differential gene expression in biofilm growth conditions relative to planktonic growth conditions was determined by Nobile et al. (2012) by reconciling information from microarray and RNA-Seq assays (Nobile et al.; 2012).

3.5.2 Variant calling and variant effect prediction

Single nucleotide variants (SNPs) and indels were identified using a custom, reproducible pipeline, as described in chapter 4. Briefly, paired-end reads were aligned to the *C. albicans* reference strain SC5314 using Bowtie2 (version 2.2.5) (Langmead and Salzberg; 2012). SNPs and indels were identified for each strain using the HaplotypeCaller function in GATK (version 4.2.0) (McKenna et al.; 2010). SVs were identified using the Genome Rearrangement IDentification Software Suite (GRIDSS) (version 2.12.0) (Cameron et al.; 2017, 2021). Simple SV events such as insertions, deletions, duplications and inversions were annotated with a custom R script using the StructuralVariantAnnotation package (version 1.10.1) (Cameron and Dong; 2021).

Effect of SNPs and indels on protein-coding genes were estimated using the Variant Effect Predictor (VEP) (version 104.3) (McLaren et al.; 2016). Effects of SVs on protein-coding genes were estimated with a custom python script (version 3.10) to annotate genes disrupted by insertions. Protein-coding genes overlapping deletions and inversions were annotated using BEDTools (version 2.30.0) (Quinlan and Hall; 2010). Effects of SNPs, indels and SVs on TFBS in the *cis*-regulatory regions were estimated using a custom pipeline. Primary and alternate motifs, representing binding preferences, were identified for the TFs. The motifs were represented as position probability matrices (PPMs), as described in chapter 2. Motif occurrences were identified in the intergenic region of all genes in the *C. albicans* reference genome using the motif PPM as input with the FIMO tool in the MEME-suite (version 5.5.2) (Grant et al.; 2011). The reference genome was updated based on the identified SNPs and indels in each strain using the FastaAlternateReferenceMaker function in GATK (version 4.2.0) (McKenna et al.; 2010). FIMO (version 5.5.2) (Grant et al.; 2011) was then run independently on each strain to obtain loss and gain of motifs for the TFs in each strain. Loss of motifs due to structural variations such as inversions and deletions were identified using BEDTools (version 2.30.0) (Quinlan and Hall; 2010). Gain of motifs due to insertions identified by retrieving inserted sequences and using FIMO (version 5.5.2) (Grant et al.; 2011), as mentioned above to identify motifs in these SVs.

3.5.3 Phylogenetic reconstruction

The SNPs obtained across all strains were used to reconstruct the phylogenetic relationships between the strains. Phylogenetic inference is sensitive to the choice of the DNA substitution model (Huelsenbeck and Crandall; 1997). ModelTest-NG (version 0.1.7) (Darriba et al.; 2020) was used to calculate the goodness-of-fit of 88 DNA substitution models to the SNP data. ModelTest-NG ranks the substitution models based on Akaike Information Criterion (AIC) (Akaike; 1974) and Bayesian Information Criterion (BIC) (Schwarz; 1978) and identified the General Time Reversible (GTR+G) model as the substitution model that best fits the data (Tavaré; 1986). The GTR model is a widely used, flexible DNA substitution model (Barba-Montoya et al.; 2020), which assumes unequal base frequencies as well as unequal substitution rates between bases (Tavaré; 1986). The GTR model assumes that the substitution rates are homogeneous across sites, so we included the discrete-gamma model (Yang; 1994) to specify variable substitution rate at each site. RAxML-NG (version 1.2.0) (Kozlov et al.; 2019) was used for phylogenetic inference using 20 starting trees for branch length estimation and 100 Felsenstein bootstraps (Felsenstein; 1985) to assess support for the branches. The phylogenetic tree estimated by RAxML-NG (Kozlov et al.; 2019) was visualized using iTOL (Letunic and Bork; 2021).

3.5.4 Detecting modes of selection

We used the SNPs identified across strains to infer the mode of selection acting on the protein-coding genes and *cis*-regulatory regions in the *C. albicans* genome. We used dN/dS to infer the role of selection in protein-coding genes and Tajima's D for *cis*-regulatory regions. We extended the alnpi module in the FAST toolkit (Lawrence et al.; 2015) to compute dN and dS from MSA, as described in chapter 4.

Tajima's D

Tajima's D is a statistic that summarizes the allele frequency spectrum in a genomic region (Tajima; 1989). The deviation of the allele frequencies from an expected frequency spectrum under assumptions of neutral evolution indicates that natural selection is acting on the genomic region of interest. Tajima's D is the normalized difference between the observed average pairwise differences between strains (π) and expected pairwise differences (θ). Tajima's D is computed as shown below:

$$\theta = \text{expectation of } \pi = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

where, S is the number of segregating sites (polymorphic sites) and n denotes the number of sequences compared. Under the null hypothesis, we expect the observed pairwise differences between strains (π) and expected pairwise differences (θ) to be equal, and the coefficient of variation $d = \pi - \theta$ would be small. Tajima's D is the coefficient of variation computed as

$$D = \frac{d}{Var(d)} = \frac{\pi - \theta}{\sqrt{e_1 S + e_2 S(S-1)}}$$

where, the coefficients are calculated as follows

$$e_1 = \frac{c_1}{a_1} \text{ and } e_2 = \frac{c_2}{a_1^2 + a_2}$$

$$c_1 = b_1 - \frac{1}{a_1} \text{ and } c_2 = b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2}$$

$$b_1 = \frac{n+1}{3(n-1)} \text{ and } b_2 = \frac{2(n^2+n+3)}{9n(n-1)}$$

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i} \text{ and } a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2}$$

Tajima's D was calculated genome-wide in 1000 bp windows and the formula for calculation was obtained from Tajima (1989).

3.5.5 Statistical tests

All statistical tests were conducted in R (version 4.1.2). For comparing normalized SNP and indel abundance in the 190 strains across gene categories, we used the Friedman test (Friedman; 1937). The Friedman test is a non-parametric test for differences between groups for dependent samples. The test assumes that the dependent variable is ordinal or continuous. The dependent variable need not be normally distributed. The Friedman χ^2 test statistic, also known as Q is given as:

$$\chi^2(Q) = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

where, n is the number of samples, k is the number of groups, R_j is the sum of ranks for the group j.

The effect size for the Friedman test is used to assess the strength of the relationship between the dependent and the independent variables. The effect size for the Friedman test (known as Kendall's W) is estimated using using the formula $W = \frac{Q}{n(k-1)}$, where Q, n and k denote the Friedman test statistic, number of samples and number of groups respectively.

For comparing the TFBS loss-to-gain ratio between TFs and the length of SVs between SV type, we used the Kruskal-Wallis nonparametric test (Kruskal and Wallis; 1952). The Kruskal-Wallis χ^2 test statistic, also known as H is given as:

$$\chi^2(H) = \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(N+1)$$

where, N is the total number of samples, k is the number of groups, R_j is the sum of ranks for sample j, n_j is the sample size of group j.

The effect size for the Kruskal-Wallis test is estimated using the formula, $H = \frac{(H-k+1)}{(N-k)}$, where H, k and N denote the Kruskal-Wallis test statistic, number of groups and total number of samples respectively.

The Kruskal-Wallis test assumes equal variance in each category. When this assumption was violated, we used the nonparametric k-sample Anderson-Darling test (Scholz and Stephens; 1987), to test if the distribution of the random variable in two or more groups are identical. The k-sample Anderson-Darling test was performed using the kSamples package (version 1.2-9) in R.

Gene Set Enrichment Analysis (GSEA)

GSEA (Subramanian et al.; 2005) is used to identify functionally enriched pathways or processes in a set of genes ranked by specific properties such as gene expression. We used genes ranked by their dN/dS estimates to identify enriched functions using the clusterProfile package (version 4.2.2) (Wu et al.; 2021) in R. We obtained gene function annotations (Gene Ontology (GO) annotations) for the *C. albicans* protein-coding genes from the Candida Genome Database (CGD) (Skrzypek et al. (2016)). GSEA determines if members of a gene set with a specific functions S tend to occur towards the top or bottom of the ranked list L. Hence, this method uses a weighted Kolmogorov-Smirnov test statistic to test whether distribution of genes in the gene set S differs from a uniform distribution. For N genes in the comparison, hits of genes g into gene set S is calculated as

$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{|r_j|}{N_R}$$

where, $N_R = \sum_{g_j \in S} |r_j|$, r_j is the ranking of gene at position j. The genes outside the gene set S are calculated using the formula below:

$$P_{\text{miss}}(S, i) = \sum_{\substack{g_j \notin S \\ j \leq i}} \frac{1}{N - N_H}$$

where, N_H is the gene set size and N is the total number of genes. The enrichment score ES is the maximum deviation of 0 from $P_{\text{hit}} - P_{\text{miss}}$. The ES value is normalized by the size of the gene set S. The statistical significance of ES is then calculated by permuting the annotation labels of the genes using a permutation test (Subramanian et al.; 2005).

3.5.6 Data visualization

The phylogenetic tree was annotated and visualized using iTOL (Letunic and Bork; 2021). Statistical tests, violin plots, density plots, plots depicting variant frequencies and correlation visualizations were performed using ggstatsplot package (version 0.11.1) in R (Patil (2021)). Results from GSEA were visualized using the clusterProfiler package (version 4.2.2) in R (Wu et al.; 2021). R version 4.1.2 was used for all R packages.

3.6 Bibliography

Afanasyeva, A., Bockwoldt, M., Cooney, C. R., Heiland, I. and Gossmann, T. I. (2018). Human long intrinsically disordered protein regions are frequent targets of positive selection, *Genome research* 28(7): 975–982.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE transactions on automatic control* 19(6): 716–723.

Anderson, M. Z., Thomson, G. J., Hirakawa, M. P. and Bennett, R. J. (2019). A 'parameiosis' drives depolyploidization and homologous recombination in Candida albicans, *Nature communications* 10(1): 4388.

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A. et al. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax, *Nature Genetics* 53(3): 354–366.

Bailey, T. L. (2021). STREME: accurate and versatile sequence motif discovery, *Bioinformatics* 37(18): 2834–2840.

Barba-Montoya, J., Tao, Q. and Kumar, S. (2020). Using a GTR+ γ substitution model for dating sequence divergence when stationarity and time-reversibility assumptions are violated, *Bioinformatics* 36(Supplement_2): i884–i894.

Bensasson, D., Dicks, J., Ludwig, J. M., Bond, C. J., Elliston, A., Roberts, I. N. and James, S. A. (2019). Diverse lineages of candida albicans live on old oaks, *Genetics* 211(1): 277–288.

Berman, J. and Hadany, L. (2012). Does stress induce (para) sex? Implications for Candida albicans evolution, *Trends in Genetics* 28(5): 197–203.

Booker, T. R., Yeaman, S. and Whitlock, M. C. (2020). Variation in recombination rate affects detection of outliers in genome scans under neutrality, *Molecular ecology* 29(22): 4274–4279.

Boytsov, A., Abramov, S., Makeev, V. J. and Kulakovskiy, I. V. (2022). Positional weight matrices have sufficient prediction power for analysis of noncoding variants, *F1000Research* 11.

Brodsky, S., Jana, T., Mittelman, K., Chapal, M., Kumar, D. K., Carmi, M. and Barkai, N. (2020). Intrinsically disordered regions direct transcription factor in vivo binding specificity, *Molecular cell* 79(3): 459–471.

Cameron, D. and Dong, R. (2021). *StructuralVariantAnnotation: Variant annotations for structural variants*. R package version 1.10.1.

Cameron, D. L., Baber, J., Shale, C., Valle-Inclan, J. E., Besselink, N., van Hoeck, A., Janssen, R., Cuppen, E., Priestley, P. and Papenfuss, A. T. (2021). GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing, *Genome biology* 22: 1–25.

Cameron, D. L., Schröder, J., Penington, J. S., Do, H., Molania, R., Dobrovic, A., Speed, T. P. and Papenfuss, A. T. (2017). GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly, *Genome research* 27(12): 2050–2060.

Cheng, C. and Kirkpatrick, M. (2021). Molecular evolution and the decline of purifying selection with age, *Nature Communications* 12(1): 2657.

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B. and Flouri, T. (2020). Modeltest-ng: a new and scalable tool for the selection of dna and protein evolutionary models, *Molecular biology and evolution* 37(1): 291–294.

Dermitzakis, E. T. and Clark, A. G. (2002). Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover, *Molecular biology and evolution* 19(7): 1114–1121.

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

Ene, I. V., Hickman, M. A. and Gerstein, A. C. (2021). The interplay between neutral and adaptive processes shapes genetic variation during Candida species evolution, *Current Clinical Microbiology Reports* 8(3): 129–138.

Feil, E. J., Cooper, J. E., Grundmann, H., Robinson, D. A., Enright, M. C., Berendt, T., Peacock, S. J., Smith, J. M., Murphy, M., Spratt, B. G. et al. (2003). How clonal is Staphylococcus aureus?, *Journal of bacteriology* 185(11): 3307–3316.

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

Fijarczyk, A. and Babik, W. (2015). Detecting balancing selection in genomes: limits and prospects, *Molecular ecology* 24(14): 3529–3545.

Fleischmann, R., Alland, D., Eisen, J. A., Carpenter, L., White, O., Peterson, J., DeBoy, R., Dodson, R., Gwinn, M., Haft, D. et al. (2002). Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains, *Journal of bacteriology* 184(19): 5479–5490.

Frazer, C., Staples, M. I., Kim, Y., Hirakawa, M., Dowell, M. A., Johnson, N. V., Hernday, A. D., Ryan, V. H., Fawzi, N. L., Finkelstein, I. J. et al. (2020). Epigenetic cell fate in candida albicans is controlled by transcription factor condensates acting at super-enhancer-like elements, *Nature microbiology* 5(11): 1374–1389.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance, *Journal of the american statistical association* 32(200): 675–701.

Gattepaille, L. M., Jakobsson, M. and Blum, M. G. (2013). Inferring population size changes with sequence and SNP data: lessons from human bottlenecks, *Heredity* 110(5): 409–419.

Gong, J., Chen, X.-F., Fan, X., Xu, J., Zhang, H., Li, R.-Y., Chen, S. C., Kong, F., Zhang, S., Sun, Z.-Y. et al. (2023). Emergence of antifungal resistant subclades in the global predominant phylogenetic population of candida albicans, *Microbiology Spectrum* 11(1): e03807–22.

Grant, C. E., Bailey, T. L. and Noble, W. S. (2011). Fimo: scanning for occurrences of a given motif, *Bioinformatics* 27(7): 1017–1018.

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.

Hirakawa, M. P., Martinez, D. A., Sakthikumar, S., Anderson, M. Z., Berlin, A., Gujja, S., Zeng, Q., Zisson, E., Wang, J. M., Greenberg, J. M. et al. (2015). Genetic and phenotypic intra-species variation in candida albicans, *Genome research* 25(3): 413–425.

Huang, M. Y., Woolford, C. A., May, G., McManus, C. J. and Mitchell, A. P. (2019). Circuit diversification in a biofilm regulatory network, *PLoS pathogens* 15(5): e1007787.

Huelsenbeck, J. P. and Crandall, K. A. (1997). Phylogeny estimation and hypothesis testing using maximum likelihood, *Annual Review of Ecology and systematics* 28(1): 437–466.

Hultqvist, G., Åberg, E., Camilloni, C., Sundell, G. N., Andersson, E., Dogan, J., Chi, C. N., Vendruscolo, M. and Jemth, P. (2017). Emergence and evolution of an interaction between intrinsically disordered proteins, *Elife* 6: e16059.

Jain, N., Kohli, R., Cook, E., Gialanella, P., Chang, T. and Fries, B. (2007). Biofilm formation by and antifungal susceptibility of candida isolates from urine, *Applied and environmental microbiology* 73(6): 1697–1703.

Johnson, M. S., Gopalakrishnan, S., Goyal, J., Dillingham, M. E., Bakerlee, C. W., Humphrey, P. T., Jagdish, T., Jerison, E. R., Kosheleva, K., Lawrence, K. R. et al. (2021). Phenotypic and molecular evolution across 10,000 generations in laboratory budding yeast populations, *Elife* 10: e63910.

Jones, T., Federspiel, N. A., Chibana, H., Dungan, J., Kalman, S., Magee, B., Newport, G., Thorstenson, Y. R., Agabian, N., Magee, P. et al. (2004). The diploid genome sequence of Candida albicans, *Proceedings of the National Academy of Sciences* 101(19): 7329–7334.

Khan, T., Douglas, G. M., Patel, P., Nguyen Ba, A. N. and Moses, A. M. (2015). Polymorphism analysis reveals reduced negative selection and elevated rate of insertions and deletions in intrinsically disordered protein regions, *Genome biology and evolution* 7(6): 1815–1826.

Korneliussen, T. S., Moltke, I., Albrechtsen, A. and Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data, *BMC bioinformatics* 14(1): 1–14.

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. and Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference, *Bioinformatics* 35(21): 4453–4455.

Krieger, G., Lupo, O., Wittkopp, P. and Barkai, N. (2022). Evolution of transcription factor binding through sequence variations and turnover of binding sites, *Genome Research* 32(6): 1099–1111.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis, *Journal of the American statistical Association* 47(260): 583–621.

Kryazhimskiy, S. and Plotkin, J. B. (2008). The population genetics of dN/dS, *PLoS genetics* 4(12): e1000304.

Kuhn, D., Chandra, J., Mukherjee, P. and Ghannoum, M. (2002). Comparison of biofilms formed by candida albicans and candida parapsilosis on bioprosthetic surfaces, *Infection and immunity* 70(2): 878–888.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2, *Nature methods* 9(4): 357–359.

Lawrence, T. J., Kauffman, K. T., Amrine, K. C., Carper, D. L., Lee, R. S., Becich, P. J., Canales, C. J. and Ardell, D. H. (2015). FAST: FAST analysis of sequences toolbox, *Frontiers in genetics* 6: 172.

Legrand, M., Jaitly, P., Feri, A., d'Enfert, C. and Sanyal, K. (2019). Candida albicans: an emerging yeast model to study eukaryotic genome plasticity, *Trends in Genetics* 35(4): 292–307.

Letunic, I. and Bork, P. (2021). Interactive tree of life (itol) v5: an online tool for phylogenetic tree display and annotation, *Nucleic acids research* 49(W1): W293–W296.

Li, X. V., Leonardi, I., Putzel, G. G., Semon, A., Fiers, W. D., Kusakabe, T., Lin, W.-Y., Gao, I. H., Doron, I., Gutierrez-Guerrero, A. et al. (2022). Immune regulation by fungal strain diversity in inflammatory bowel disease, *Nature* 603(7902): 672–678.

Li, X., Yan, Z. and Xu, J. (2003). Quantitative variation of biofilms among strains in natural populations of candida albicans, *Microbiology* 149(2): 353–362.

Lott, T. J., Holloway, B. P., Logan, D. A., Fundyga, R. and Arnold, J. (1999). Towards understanding the evolution of the human commensal yeast candida albicans, *Microbiology* 145(5): 1137–1143.

Ludwig, M. Z., Bergman, C., Patel, N. H. and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element, *Nature* 403(6769): 564–567.

MacCallum, D. M., Castillo, L., Nather, K., Munro, C. A., Brown, A. J., Gow, N. A. and Odds, F. C. (2009). Property differences among the four major Candida albicans strain clades, *Eukaryotic cell* 8(3): 373–387.

Mancera, E., Nocedal, I., Hammel, S., Gulati, M., Mitchell, K. F., Andes, D. R., Nobile, C. J., Butler, G. and Johnson, A. D. (2021). Evolution of the complex transcription network controlling biofilm formation in candida species, *Elife* 10: e64682.

Mao, A. H., Crick, S. L., Vitalis, A., Chicoine, C. L. and Pappu, R. V. (2010). Net charge per residue modulates conformational ensembles of intrinsically disordered proteins, *Proceedings of the National Academy of Sciences* 107(18): 8183–8188.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data, *Genome research* 20(9): 1297–1303.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P. and Cunningham, F. (2016). The ensembl variant effect predictor, *Genome biology* 17(1): 1–14.

McManus, B. A. and Coleman, D. C. (2014). Molecular epidemiology, phylogeny and evolution of Candida albicans, *Infection, Genetics and Evolution* 21: 166–178.

Moses, A. M., Pollard, D. A., Nix, D. A., Iyer, V. N., Li, X.-Y., Biggin, M. D. and Eisen, M. B. (2006). Large-scale turnover of functional transcription factor binding sites in Drosophila, *PLoS computational biology* 2(10): e130.

Necci, M., Piovesan, D., Clementel, D., Dosztányi, Z. and Tosatto, S. C. (2020). MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavors in proteins, *Bioinformatics* 36(22-23): 5533–5534.

Nilsson, J., Grahn, M. and Wright, A. P. (2011). Proteome-wide evidence for enhanced positive darwinian selection within intrinsically disordered regions in proteins, *Genome biology* 12(7): 1–17.

Nobile, C. J., Fox, E. P., Nett, J. E., Sorrells, T. R., Mitrovich, Q. M., Hernday, A. D., Tuch, B. B., Andes, D. R. and Johnson, A. D. (2012). A recently evolved transcriptional network controls biofilm development in Candida albicans, *Cell* 148(1-2): 126–138.

Nystrom, S. L. and McKay, D. J. (2021). Memes: A motif analysis environment in R using tools from the MEME Suite, *PLoS Computational Biology* 17(9): e1008991.

Odds, F. C., Bougnoux, M.-E., Shaw, D. J., Bain, J. M., Davidson, A. D., Diogo, D., Jacobsen, M. D., Lecomte, M., Li, S.-Y., Tavanti, A. et al. (2007). Molecular phylogenetics of Candida albicans, *Eukaryotic cell* 6(6): 1041–1052.

Oldfield, C. J. and Dunker, A. K. (2014). Intrinsically disordered proteins and intrinsically disordered protein regions, *Annual review of biochemistry* 83: 553–584.

Patil, I. (2021). Visualizations with statistical details: The'ggstatsplot'approach, *Journal of Open Source Software* 6(61): 3167.

Piovesan, D., Necci, M., Escobedo, N., Monzon, A. M., Hatos, A., Mičetić, I., Quaglia, F., Paladin, L., Ramasamy, P., Dosztányi, Z. et al. (2021). MobiDB: intrinsically disordered proteins in 2021, *Nucleic acids research* 49(D1): D361–D367.

Pujol, C., Daniels, K. J. and Soll, D. R. (2015). Comparison of switching and biofilm formation between mtl-homozygous strains of candida albicans and candida dubliniensis, *Eukaryotic Cell* 14(12): 1186–1202.

Quinlan, A. R. and Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features, *Bioinformatics* 26(6): 841–842.

Radman, M., Matic, I. and Taddei, F. (1999). Evolution of evolvability a, *Annals of the New York Academy of Sciences* 870(1): 146–155.

Ridout, K. E., Dixon, C. J. and Filatov, D. A. (2010). Positive selection differs between protein secondary structure elements in drosophila, *Genome biology and evolution* 2: 166–179.

Romeo, O. and Criseo, G. (2011). Candida africana and its closest relatives, *Mycoses* 54(6): 475–486.

Ropars, J., Maufrais, C., Diogo, D., Marcet-Houben, M., Perin, A., Sertour, N., Mosca, K., Permal, E., Laval, G., Bouchier, C. et al. (2018). Gene flow contributes to diversification of the major fungal pathogen candida albicans, *Nature communications* 9(1): 2253.

Schnepf, M., von Reutern, M., Ludwig, C., Jung, C. and Gaul, U. (2020). Transcription factor binding affinities and DNA shape readout, *Iscience* 23(11).

Scholz, F. W. and Stephens, M. A. (1987). K-sample Anderson–Darling tests, *Journal of the American Statistical Association* 82(399): 918–924.

Schwarz, G. (1978). Estimating the dimension of a model, *The annals of statistics* pp. 461–464.

Selmecki, A., Forche, A. and Berman, J. (2010). Genomic plasticity of the human fungal pathogen Candida albicans, *Eukaryotic cell* 9(7): 991–1008.

Sigrist, C. J., De Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., Bougueleret, L. and Xenarios, I. (2012). New and continuing developments at PROSITE, *Nucleic acids research* 41(D1): D344–D347.

Sitterlé, E., Maufrais, C., Sertour, N., Palayret, M., d'Enfert, C. and Bougnoux, M.-E. (2019). Within-host genomic diversity of candida albicans in healthy carriers, *Scientific reports* 9(1): 1–12.

Skrzypek, M. S., Binkley, J., Binkley, G., Miyasato, S. R., Simison, M. and Sherlock, G. (2016). The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data, *Nucleic acids research* p. gkw924.

Soll, D. R. and Daniels, K. J. (2016). Plasticity of candida albicans biofilms, *Microbiology and molecular biology reviews* 80(3): 565–595.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Sciences* 102(43): 15545–15550.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism., *Genetics* 123(3): 585–595.

Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequence., *Lecture of Mathematics for Life Science* 17: 57.

The UniProt Consortium (2023). UniProt: the universal protein knowledgebase in 2023, *Nucleic Acids Research* 51(D1): D523–D531.

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

Todd, R. T., Wikoff, T. D., Forche, A. and Selmecki, A. (2019). Genome plasticity in Candida albicans is driven by long repeat sequences, *Elife* 8: e45954.

Villar-Vidal, M., Marcos-Arias, C., Eraso, E. and Quindós, G. (2011). Variation in biofilm formation among blood and oral isolates of candida albicans and candida dubliniensis, *Enfermedades Infecciosas y Microbiologia Clinica* 29(9): 660–665.

Wang, J. M., Bennett, R. J. and Anderson, M. Z. (2018). The genome of the human pathogen Candida albicans is shaped by mutation and cryptic sexual recombination, *MBio* 9(5): 10–1128.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L. et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data, *The Innovation* 2(3): 100141.

Yang, Z. (1994). Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods, *Journal of Molecular evolution* 39: 306–314.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. et al. (2008). Model-based analysis of ChIP-Seq (MACS), *Genome biology* 9(9): 1–9.

Chapter 4

Bioinformatic tools for identifying transcription factor binding sites, variants within species and calculating selection statistics

4.1 Introduction

This chapter is comprised of three bioinformatic tools that I developed for the work presented in chapter 2 and chapter 3. These tools serve as standalone workflows that can be used to process high-throughput, next-generation sequenced data. The first workflow was developed for end-to-end processing of ChIP-Seq data, from read alignment to binding site annotation and motif calling. This ChIP-Seq workflow was developed to improve sensitivity in identifying binding sites for TFs, especially for single-end and short-reads (approximately, 50 bps). The second workflow was developed for variant calling, integrating identification of SNPs and SVs in microbial species. This variant calling workflow is tailored for variant identification in non-model organisms without benchmarked, known variants, which are traditionally used to distinguish sequencing errors from real variants. In the absence of these benchmark datasets, we leverage multiple rounds of error correction to increase precision of SNP identification. Furthermore, a variant effect prediction function was also developed to predict the effect of variants on binding site loss and gain in *cis*-regulatory regions for known motifs. The manuscript for this workflow is currently in preparation for submission to *Current Protocols in Bioinformatics*. The last section of this chapter is dedicated to the expansion of a previously published module alnpi (Lawrence et al.; 2015). This module was published as part of the FAST toolkit (Lawrence et al.; 2015) and is used to compute commonly used molecular evolution statistics from multiple sequence alignments MSA.

4.2 Identifying transcription factor binding sites and motifs

This workflow was developed to identify transcription factor binding sites (TFBSs) from single-end reads generated from ChIP-Seq experiments. ChIP-Seq experiments are used to determine genome-wide transcription factor (TF) binding events. The general approach of a ChIP-Seq experiment is depicted in Figure 4.1. The genomic loci bound by the TFs are purified, amplified and sequenced. The reads are then mapped to the reference genome, and binding peaks are identified using the workflow described below. This module is available for download at (https://github.com/dgunasekaran/chipper_motif.git).

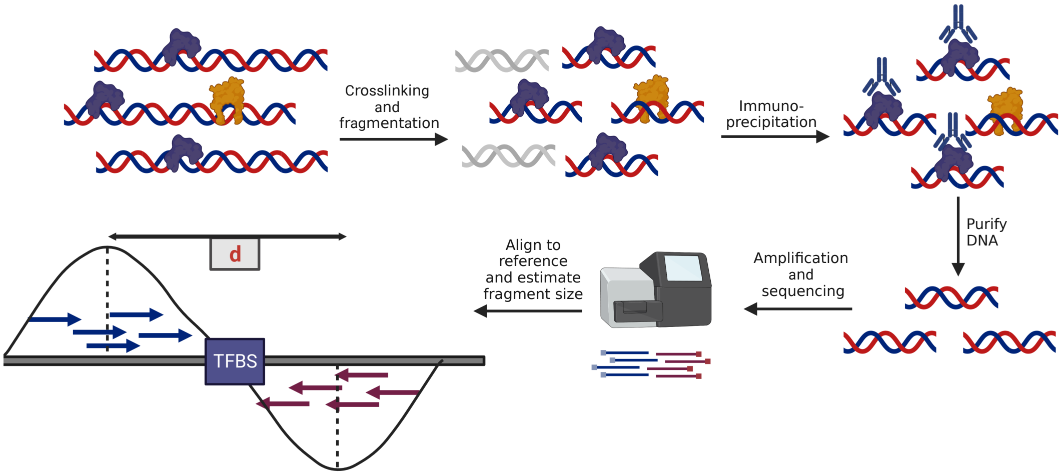


Figure 4.1: **ChIP-Seq overview.**
General approach of a ChIP-Seq experiment. TFs (DNA-binding proteins) bind to genomic DNA and ChIP-Seq is used to determine the binding loci of the TFs. First, the protein-DNA interaction is stabilized by cross-linking. The DNA is then sheared to obtain smaller fragments. The samples are then treated with an antibody that interacts with the TF of interest, which is then used to extract binding loci specific to the corresponding TF. The TF is then decoupled from the genomic DNA and the binding loci of interest are purified, amplified and sequenced. The reads are sequenced from the 5'-ends, therefore, mapping of the reads to the reference genomes gives us a set of reads mapping to the sense (Watson) strand and a set of reads mapping to the anti-sense (Crick) strand. The peaks of reads in the sense and anti-sense strands are separated by a distance d and the actual binding site is located at approximately $d/2$. BioRender was used to make this figure.

4.2.1 Data source

The workflow for processing ChIP-Seq data was developed using the Snakemake workflow management systems (version 6.10.0) (Mölder et al.; 2021) for reproducible data analyses. This workflow management system also helps with optimum utilization of compute resources, especially for analyzing high-throughput (epi-)genomic data. This workflow was used to analyze data from Mancera et al. (2021) used in chapter 2. ChIP-Seq binding data was obtained from NCBI Gene Expression Omnibus (GEO) repository. We obtained the unphased reference genome and annotation for *C. albicans* from the Candida Genome Database (Skrzypek et al.; 2016). *C. dubliniensis*, *C. tropicalis* and *C. parapsilosis* genomes were obtained from NCBI. We selected the genomes for the same strains as the ones used for ChIP-Seq experiments by Mancera et al. (2021). To facilitate comparison of binding events across species, orthologs of protein-coding genes were identified across the four species as described in chapter 2. Briefly, orthologs and paralogs of *C. albicans* genes were identified across species using OrthoFinder (version 2.5.4) (Emms and Kelly (2015, 2019)). OrthoFinder was run with parameters `-S diamond` to identify sequence similarity using DIAMOND (Buchfink et al. (2015)) and `-M msa` to perform multiple sequence alignment (MSA) of orthogroups with MAFFT (Katoh and Standley (2013)) and infer gene trees from MSA using FastTree (Price et al. (2010)). Although we used the above-mentioned dataset for this study, the pipeline was developed to handle any single-end ChIP-seq data, with corresponding reference genome (in FASTA format) and protein-coding gene annotations (in GFF format) as inputs.

4.2.2 Data quality control and read mapping

Quality of raw, single-end reads, sequenced using Illumina HiSeq 2500 or 4000 platforms was verified using FastQC (version 0.11.9) (Andrews et al.; 2010). Bowtie2 (version 2.2.5) was used to map the reads to the indexed reference genome (Langmead and Salzberg; 2012). Bowtie2 uses a seed-and-extend based alignment, where the reads are split into "seeds" of length 20, which are aligned to the reference without permitting gaps or mismatches. Seeds are weighted based on the number of query sequences they bind to, where unique seed are weighted higher. To ensure adapter sequences are removed, a local alignment was performed (`--local` parameter) to soft-clip parts of reads that have low quality scores. Reads that map to multiple loci in the genome (multi-mapping reads) were randomly assigned to a locus (using the `--non-deterministic` flag). Only the best match of the read to the reference (also known as primary reads) are retained using samtools (version 1.15.1) (Li et al.; 2009; Danecek et al.; 2021).

4.2.3 Identifying and annotating binding sites

Binding sites of TFs are identified by separating signals in immunoprecipitated (IP) samples (Figure 4.1) from background noise. An important parameter used for peak calling is the average fragment size (also known as shift size parameter). As shown in Figure 4.1, the reads obtained from ChIP-Seq experiments are sequenced from the 5'-ends, therefore, mapping of the reads to the reference genomes results in a peak in the sense strand and another in the anti-sense strand. These peaks are separated by a distance d and the actual binding site is located at approximately $d/2$. This estimated distance d is the average fragment size. We estimate d using a custom R script (version 4.1.2) with the package csaw (version 1.28.0) (Lun and Smyth; 2014, 2016). The reads on the sense and anti-sense strands were shifted by 10 base pairs (bps) using a sliding window approach and the cross-correlation was measured for each of the distances. The fragment distance with the maximum cross-correlation value is estimated to be the average fragment size. When binding loci are enriched in repeat regions, this can result in a bias in estimation of the average fragment size, since this results in the size corresponding to the read length having the highest cross-correlation value. To overcome this bias, we deduplicate the reads to estimate the average fragment size. Using deduplicated data for peak calling, however, reduces sensitivity. Therefore, we retain duplicate reads for subsequent steps.

We then identify the binding sites using the peak caller, MACS2 (version 2.2.7.1) (Zhang et al.; 2008). MACS2 uses the average fragment size, the effective genome size and the control sample (without the immunoprecipitation step) to identify peaks in the IP samples. The control provides an estimate of the background noise, since the genome-wide binding signal is highly variable and very sensitive to local context such as chromatin structure and accessibility (Ramachandran et al.; 2015). To handle duplicated reads, we set the parameter `--keep-dup auto`. When this parameter is specified, MACS2 estimates the maximum number of reads to use in a genomic location based on a binomial distribution (with a p-value threshold of $1e-5$). If more than the expected number of duplicated reads are identified in the locus, they are discarded. The distribution of reads along the genome is then estimated using a Poisson distribution, with λ denoting the expected number of reads in a window.

$$P_{\lambda}(X = k) = \frac{\lambda^k}{k! * e^{-\lambda}}$$

$$\lambda = \frac{\text{Read length} * \text{number of reads}}{\text{Effective genome size}}$$

MACS2 estimates λ dynamically in regions of 1 kb, 5 kb and 10 kb regions in addition to genome-wide. A region in the IP sample is considered significantly enriched if the p-value, computed as shown above and corrected for multiple comparisons using the Benjamini-Hochberg correction (Benjamini and Hochberg; 1995), is less than 0.01. The identified binding sites are then annotated using a custom python script (version

3.10). The annotation generates an output with the protein-coding genes immediately downstream of the binding sites as well as an additional file summarizing the number, strength and location of TF binding events for each gene in the genome.

4.2.4 Motif finding

The binding peaks identified using MACS2 were sorted based on the enrichment of signal in the IP sample compared to the control. The summits of the binding peaks were extended to 250 bp on either side and this 500 bp sequence was used as the input to identify motifs using STREME from the MEME-suite (version 5.5.2) (Bailey (2021)) and the memes package (version 1.2.5) in R (Nystrom and McKay (2021)). The minimum allowed motif width was 6 and the maximum was 15. A supplementary python script generates a mutational landscape for a motif (as shown in Figure 2.17) with the position probability matrix of the motif as an input. This script uses python version 3.10 and the python packages scipy (version 1.10.0), matplotlib (version 3.6.2) and numpy (version 1.23.5).

4.3 Identifying SNPs and SVs in *C. albicans*

This workflow was developed to identify SNPs and SVs from short-read paired-end data. The workflow was developed using the Snakemake workflow management systems (version 6.10.0) (Mölder et al.; 2021) for reproducible data analyses. Optimizing compute resources is crucial for this workflow, especially when working with multiple strains. In addition to variant calling, custom scripts are used to annotate structural variants and variants in *cis*-regulatory regions. Although, we used this pipeline for *C. albicans*, this workflow was developed for identifying SNPs and SVs in non-model organisms without known, experimentally verified variant data. The steps in this pipeline is depicted in Figure 4.2 and this module is available for download at (https://github.com/dgunasekaran/snp_svant.git).

4.3.1 Data acquisition, quality control and read mapping

This workflow takes as input sequence read archive (SRA) identifiers. Paired-end, short-read, genomic DNA sequences are retrieved from NCBI using the SRA toolkit (version 3.0.3). The quality of raw data are verified using FastQC (version 0.11.9) (Andrews et al.; 2010). Bowtie2 (version 2.2.5) is used to map the reads to the indexed reference genome (Langmead and Salzberg; 2012) as described in the previous section with `--local` and `--non-deterministic` parameters. Aligned reads are sorted by

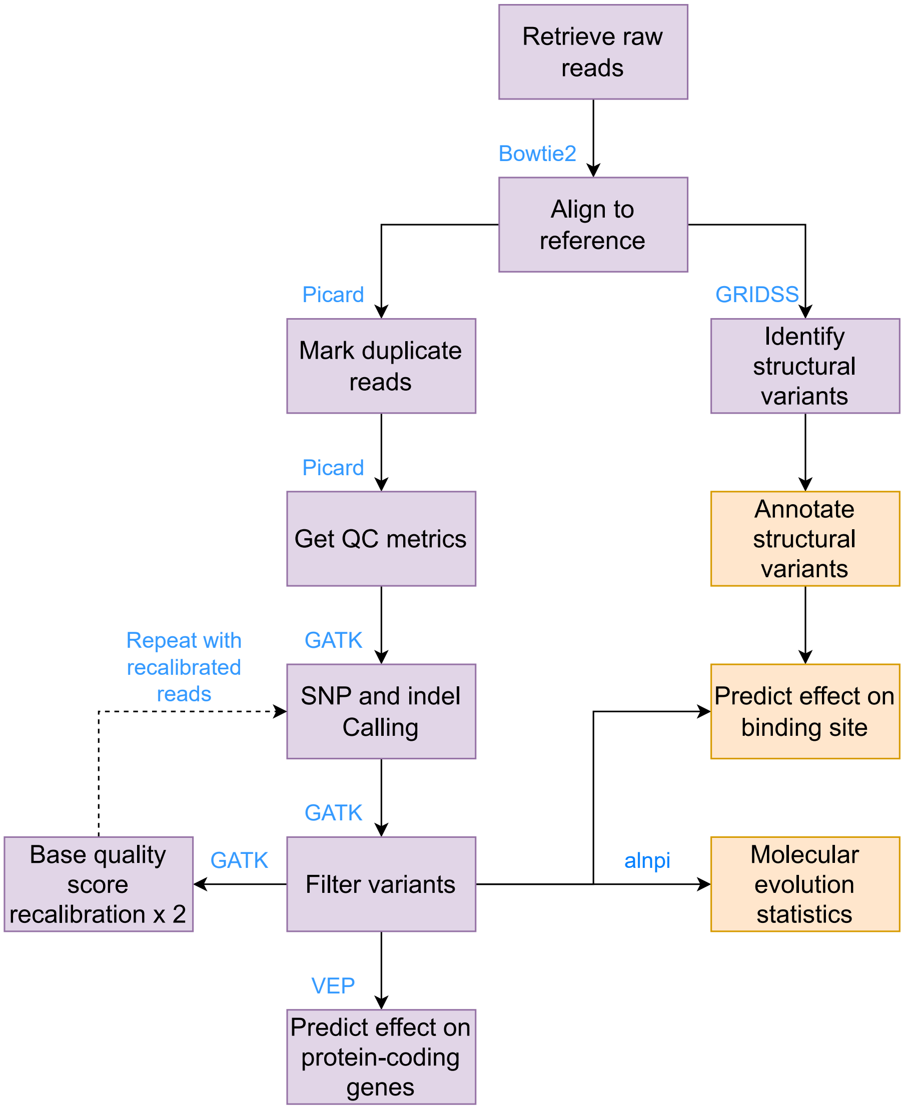

Retrieve raw reads
Bowtie2
Align to reference
Picard
Mark duplicate reads
GRIDSS
Identify structural variants
Picard
Get QC metrics
Annotate structural variants
Repeat with recalibrated reads
GATK
SNP and indel Calling
Predict effect on binding site
GATK
Base quality score recalibration x 2
GATK
Filter variants
alnpi
Molecular evolution statistics
VEP
Predict effect on protein-coding genes

Figure 4.2: **Variant calling workflow.**
We developed a reproducible, scalable variant calling pipeline to identify SNPs, indels and SVs in non-model organisms. The boxes in purple indicate steps using published tools and the boxes in yellow depict modules that we developed.

genomic loci using samtools (version 1.15.1) (Li et al.; 2009; Danecek et al.; 2021).

4.3.2 Identifying SNPs and Indels

We then mark duplicated reads in the alignment using the MarkDuplicates command in the Picard Toolkit (version 2.25.4) (*Picard toolkit*; 2019). In this step, reads duplicated during the PCR amplification step of sequencing are marked to ensure that PCR duplicated reads are excluded from calculation of statistics to assess variant quality. Summary of read alignment statistics, insert size distribution, and read depth, are also generated to verify quality of sequenced data and alignment, using the Picard Toolkit (version 2.25.4) (*Picard toolkit*; 2019) and samtools (version 1.15.1) (Li et al.; 2009; Danecek et al.; 2021). The HaplotypeCaller function in GATK (version 4.2.0) (McKenna et al.; 2010) is used to perform the first round of variant calling to identify SNPs and short indels. The identified SNPs are retained if they meet all the criteria listed below:

1. Quality score of the variant normalized by depth > 2.

2. Phred-scaled probability that the site has a strand bias < 60, where the Phred score $Q = -logP(strandbias)$. The probability of strand bias is calculated using Fisher's exact test (Fisher; 1970).

3. The strand odds ratio is another metric that estimates the strand bias, especially for regions with high coverage. Variant with strand odds ratio < 4 are retained.

4. Root mean squared mapping quality over all the reads at the site of the variant > 40.

5. Approximation from the rank sum test (Mann and Whitney; 1947) for mapping qualities of reads supporting the reference allele and reads supporting the alternate allele > -12.5.

6. Approximation from the rank sum test (Mann and Whitney; 1947) for the variant site position within the reads > -8.

The identified indels are retained if they meet all the criteria listed below:

1. Quality score of the variant normalized by depth > 2.

2. Phred-scaled probability that the site has a strand bias < 200. The probability of strand bias is calculated using Fisher's exact test (Fisher; 1970).

3. The strand odds ratio < 10.

The above thresholds are set according to GATK best practices based on comparing these metrics between true, known variants in humans and variants called using the HaplotypeCaller. These thresholds can be modified by users based on the distributions of these metrics in their data. The variants that did not meet any of the above criteria are filtered out. Base quality scores of the aligned reads are then recalibrated using the filtered, high-quality variants with the BaseQualityScoreRecalibration step. This step accounts for non-random errors during sequencing that can arise from sequence context and position of the base in the read and recalibrates the quality score of the base. This step is repeated twice and the aligned reads with the recalibrated base scores are used to perform another round of variant calling using HaplotypeCaller followed by filtering using the above criteria. These filtered variants following the second round of variant calling are retained as the final, high-confidence variants. The effect of these variants on protein coding regions is predicted using the Variant Effect Predictor (VEP) (version 104.3) (McLaren et al.; 2016)

4.3.3 Identifying and annotating structural variations

SVs can be identified from short paired-end reads based on the distance and orientation of the read pairs. Different types of structural variations result in different patterns of mapping of the read pair (Mahmoud et al.; 2019). The Genome Rearrangement IDentification Software Suite (GRIDSS) (version 2.12.0) (Cameron et al.; 2017, 2021) uses short paired-end reads to identify SVs. GRIDSS retrieves partially aligned reads, reads with incorrect orientation, reads with only one mapped end and reads with unusual insert sizes and performs further filtering based on mapping quality and complexity in mapped bases. These reads are then used to construct a positional de Bruijn graph which is used to identify break-ends iteratively. The break-end contigs are aligned to the reference and all breakpoints are identified. GRIDSS reports only the break-ends and does not annotate the variants. Hence, a custom R (version 4.1.2) script was used to annotate SVs using the StructuralVariantAnnotation package (version 1.10.1) (Cameron and Dong; 2021). During the annotation step, we further filter the identified SVs based on the reported break-ends. If the break-ends are not paired, these SVs are removed from final annotation. Since GRIDSS is optimized for sensitivity, we perform this filtering to reduce false-positive SVs.

4.4 alnpi 2.0: Calculating selection statistics from multiple sequence alignments

alnpi was originally published as part of the FAST toolkit to compute selection statistics from multiple-sequence alignments (Lawrence et al.; 2015). In this study, we expanded the alnpi module (referred to as alnpi 2.0) to include computation of selection statistics in a pairwise manner between sequences and included computation of dN/dS, a codon-based metric to estimate selection in protein-coding genes. This updated module is written in Perl and was executed with Perl version 5.30.3. This module is available for download at (https://github.com/dgunasekaran/alnpi_2_0.git).

4.4.1 Previous work

alnpi was developed by Lawrence et al. (2015). alnpi takes as input a multiple sequence alignment and returns as output the following statistics:

1. Heterozygosity.
2. Expected number of alleles (Ewens; 1972).
3. Watterson's θ (Watterson; 1975), to estimate the genetic diversity in a population with no recombination. The standard error of the estimate is also reported under assumption of no recombination, or free recombination.
4. Nucleotide diversity (π) (Nei and Li; 1979) and its standard error under assumption of no recombination, or free recombination.
5. Tajima's D (Tajima; 1989), to detect departures of allele frequency from neutral expectations based on estimates of number of segregating sites and mean pairwise difference between sequences.
6. Fu and Li's D* (Fu and Li; 1993), to detect recent selection events by comparing singleton mutations to total number of derived variants.
7. Fu and Li's F* (Fu and Li; 1993; Simonsen et al.; 1995), to detect selection events and demographic changes by comparing singleton mutations to the average number of nucleotide differences.

I expanded this module to include estimation of synonymous and nonsynonymous substitution rates to detect selection events in protein coding genes.

4.4.2 Codon analysis using dN/dS in protein-coding genes

The dN/dS ratio is one of the most widely used measures to detect evolutionary pressures acting on protein-coding genes. This measure is the ratio of nonsynonymous substitutions to synonymous substitutions, where $dN/dS > 1$ is indicative of positive selection $dN/dS < 1$ is indicative of negative selection and $dN/dS = 1$ indicates neutrality. We implemented the dN and dS estimation using the Nei and Gojobori method (Nei and Gojobori; 1986). We first calculate the number of possible synonymous (s) and nonsynonymous (n) sites, assuming equal probabilities of all nucleotide changes.

$$s = \sum_{i=1}^{3} f_i$$

where i denoted the i-th nucleotide of the codon and f_i is the ratio of synonymous changes to the sum of synonymous and nonsynonymous changes at nucleotide position i. The number of possible nonsynonymous sites is given by $n = 3 - s$. Total number of possible synonymous (S) sites and nonsynonymous (N) sites across the sequence are obtained using the formulae, $S = \sum_{j=1}^{C} s_j$ and $N = 3C - S$, respectively. C is the total number of codons in the sequence and s_j is the value of s for the j-th codon.

After obtaining S and N, we calculate the number of observed synonymous S_d and nonsynonymous N_d changes. For each codon pair, s_d and n_d are calculated. If more that one position of the codon is different between the sequences, all mutational paths between the codons are considered, and s_d is a fraction of the number of synonymous changes in all paths over the number of mutational paths. Therefore, $S_d = \sum_{j=1}^{C} s_{dj}$ and $N_d = \sum_{j=1}^{C} n_{dj}$. The proportion of synonymous and nonsynonymous substitutions are estimated using $p_S = S_d/S$ and $p_N = N_d/N$, respectively. These proportions are corrected to account for multiple substitutions using the Jukes-Cantor correction (Jukes et al.; 1969) and the number of synonymous and nonsynonymous substitutions per site are estimated using the formulae below:

$$d_S = -\frac{3}{4}ln(1-\frac{4}{3}p_S)$$
$$d_N = -\frac{3}{4}ln(1-\frac{4}{3}p_N)$$

The variance of these estimated are also calculated using:

$$Var(d_S) = \frac{p_S(1-p_S)}{(1-\frac{4}{3}p_S)^2 S}$$
$$Var(d_N) = \frac{p_N(1-p_N)}{(1-\frac{4}{3}p_N)^2 N}$$

4.5 Bibliography

Andrews, S. et al. (2010). FastQC: a quality control tool for high throughput sequence data.

Bailey, T. L. (2021). STREME: accurate and versatile sequence motif discovery, *Bioinformatics* 37(18): 2834–2840.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal statistical society: series B (Methodological)* 57(1): 289–300.

Buchfink, B., Xie, C. and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND, *Nature methods* 12(1): 59–60.

Cameron, D. and Dong, R. (2021). *StructuralVariantAnnotation: Variant annotations for structural variants.* R package version 1.10.1.

Cameron, D. L., Baber, J., Shale, C., Valle-Inclan, J. E., Besselink, N., van Hoeck, A., Janssen, R., Cuppen, E., Priestley, P. and Papenfuss, A. T. (2021). GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing, *Genome biology* 22: 1–25.

Cameron, D. L., Schröder, J., Penington, J. S., Do, H., Molania, R., Dobrovic, A., Speed, T. P. and Papenfuss, A. T. (2017). GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly, *Genome research* 27(12): 2050–2060.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M. et al. (2021). Twelve years of SAMtools and BCFtools, *Gigascience* 10(2): giab008.

Emms, D. M. and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy, *Genome biology* 16(1): 1–14.

Emms, D. M. and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics, *Genome biology* 20: 1–14.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles, *Theoretical population biology* 3(1): 87–112.

Fisher, R. A. (1970). Statistical methods for research workers, *Breakthroughs in statistics: Methodology and distribution*, Springer, pp. 66–70.

Fu, Y.-X. and Li, W.-H. (1993). Statistical tests of neutrality of mutations., *Genetics* 133(3): 693–709.

Jukes, T. H., Cantor, C. R. et al. (1969). Evolution of protein molecules, *Mammalian protein metabolism* 3: 21–132.

Katoh, K. and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability, *Molecular biology and evolution* 30(4): 772–780.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2, *Nature methods* 9(4): 357–359.

Lawrence, T. J., Kauffman, K. T., Amrine, K. C., Carper, D. L., Lee, R. S., Becich, P. J., Canales, C. J. and Ardell, D. H. (2015). FAST: FAST analysis of sequences toolbox, *Frontiers in genetics* 6: 172.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Subgroup, . G. P. D. P. (2009). The sequence alignment/map format and SAMtools, *bioinformatics* 25(16): 2078–2079.

Lun, A. T. and Smyth, G. K. (2014). De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly, *Nucleic acids research* 42(11): e95–e95.

Lun, A. T. and Smyth, G. K. (2016). csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows, *Nucleic acids research* 44(5): e45–e45.

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C. and Sedlazeck, F. J. (2019). Structural variant calling: the long and the short of it, *Genome biology* 20(1): 1–14.

Mancera, E., Nocedal, I., Hammel, S., Gulati, M., Mitchell, K. F., Andes, D. R., Nobile, C. J., Butler, G. and Johnson, A. D. (2021). Evolution of the complex transcription network controlling biofilm formation in Candida species, *Elife* 10: e64682.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other, *The annals of mathematical statistics* pp. 50–60.

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

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P. and Cunningham, F. (2016). The ensembl variant effect predictor, *Genome biology* 17(1): 1–14.

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A. et al. (2021). Sustainable data analysis with Snakemake, *F1000Research* 10.

Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions., *Molecular biology and evolution* 3(5): 418–426.

Nei, M. and Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases., *Proceedings of the National Academy of Sciences* 76(10): 5269–5273.

Nystrom, S. L. and McKay, D. J. (2021). Memes: A motif analysis environment in R using tools from the MEME suite, *PLoS Computational Biology* 17(9): e1008991.

Picard toolkit (2019). `http://broadinstitute.github.io/picard/`.

Price, M. N., Dehal, P. S. and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments, *PloS one* 5(3): e9490.

Ramachandran, P., Palidwor, G. A. and Perkins, T. J. (2015). BIDCHIPS: bias decomposition and removal from ChIP-seq data clarifies true binding signal and its functional correlates, *Epigenetics & chromatin* 8: 1–16.

Simonsen, K. L., Churchill, G. A. and Aquadro, C. F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data., *Genetics* 141(1): 413–429.

Skrzypek, M. S., Binkley, J., Binkley, G., Miyasato, S. R., Simison, M. and Sherlock, G. (2016). The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data, *Nucleic acids research* p. gkw924.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism., *Genetics* 123(3): 585–595.

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination, *Theoretical population biology* 7(2): 256–276.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W. et al. (2008). Model-based analysis of ChIP-Seq (MACS), *Genome biology* 9(9): 1–9.

Chapter 5

Conclusion

5.1 Scientific Impact

In this study, we used a well-defined gene regulatory network (GRN) in *Candida albicans* to identify sources of phenotypic diversity within species and divergence of this network across species. We found that the network structure has diverged between *Candida* species and diversity in *C. albicans* strains is driven by genomic variants in transcriptional regulators and loss of binding sites of these regulators. We developed three tools to analyze genomic and epigenomic data and to compute population genomic statistics. Tools that identify genomic variants and predict effects of variants in protein-coding genes as well as *cis*-regulatory regions are available for well studied species, such as humans, with an abundance of population genomic data. In this work, we optimized these tools for use with non-model organisms that do not have benchmarked, ground-truth datasets. Using these tools, we identified novel functional targets of the biofilm GRN in *C. albicans.* Most studies exploring genomic variation of *cis*-regulatory regions have focused on SNPs and genomic variants that affect individual transcription factor binding sites. Here, we also identified the effect of larger structural variations SVs on *cis*-regulatory regions using the variant calling tool we developed. The tool we developed for identifying binding sites and motifs was optimized for sensitivity and this helped us understand the divergence in regulatory network structures across different *Candida* species. This tools also uses ortholog information across multiple species to annotate divergence in target genes in the network. By comparing the motif preferences of transcription factors TFs across species, we were able to visualize the mutational landscape of the motifs based on the estimated binding strength of mutational neighbors.

Using these tools we answered broad questions about the divergence and evolution of the biofilm regulatory network. We found that interspecific divergence of the biofilm regulatory network is driven primarily by altering the network structure to regulate genes in a concerted manner. This is consistent with the enhanceosome model as opposed to the billboard model of regulation (Arnosti and Kulkarni; 2005). These models posit different theories on the roles of multiple binding sites found in *cis*-regulatory regions. According to the enhanceosome model (also known as the regulatory grammar model (Weingarten-Gabbay and Segal; 2014)), transcription factors are highly cooperative, binding to the *cis*-regulatory regions and altering gene expression of their downstream genes in a concerted manner. This cooperative binding is highly dependent on orientation, ordering and spacing of binding sites. This implies that individual binding sites are evolutionarily constrained since mutations that affect binding of even a single transcription factor affect the assembly of the complex that drives gene expression. In contrast, the billboard model suggests that binding sites are flexible and the the relative orientation of binding sites are less relevant to transcriptional regulation. According to this model, individual binding sites in *cis*-regulatory regions are less conserved, relying instead on the additive recruitment of transcription factors. This indicates that the number of binding sites in a regulatory region are conserved. In our study, we see that binding site loss is prevalent but the frequency of indels and SNPs are lower in regulatory regions of the biofilm network. Additionally, we also show that intrinsically disordered regions in protein-coding genes, regions that are essential for recruiting other regulators to the promoter, frequently accumulate mutations and are likely under weaker purifying selection or balancing selection. Regulators with intrinsically disordered regions can recruit other transcription factors to the *cis*-regulatory region, which can bind to lower affinity or non-cognate motifs (Staller; 2022). This implies that motif spacing between regulators is more constrained than motif strength. This theory fits our observation of prevalent binding site losses and gains of lower affinity binding sites.

We also found different sources of genetic variation in the biofilm network components of *C. albicans*. Large structural variants are widely found in *C. albicans* strains. A recent study found frequent structural variations in commensal strains even within the same host (Anderson et al.; 2023). Hence, structural variations contribute significantly to genetic diversity in *C. albicans*. When we examined the effect of these structural variations on binding sites, we found that losses of binding sites due to large deletions are compensated by binding site gains in duplications. Although insertions were also found in *C. albicans* strains, the insertion sizes are small and the binding site gains in these insertions do not compensate completely for binding site losses due to deletions. Although binding site losses are compensated at a genome-wide scale, individual regulatory regions are more likely subjected to dosage effects caused by structural variations (Chiang et al.; 2017). The variant allele frequencies of structural variations, however, are typically low (Sudmant et al.; 2015), hence, we expect this dosage effect to be minimal.

We estimated the empirical rates of losses and gains of binding sites for each transcription factor in the biofilm regulatory network. We also found some functional gene categories with more binding site gains than losses. These include, for example, the GPI-anchored cell wall adhesins, a family of proteins that has expanded in *C. albicans* (Nather and Munro; 2008) and is important for host-pathogen interactions (Martin et al.; 2021). We also found that GPI-anchored proteins are recruited to the *C. albicans* biofilm regulatory network by multiple transcription factors, possibly through binding site gains.

5.2 Limitations

In this study, we assessed the conservation of biofilm regulators and target genes across species. We identified that orthologous relationships for some genes (for example, *EFG1* and *FLO8*) were mispredicted in some species due to gaps in genome assembly or due to missense variants unique to the reference strain of the species. Even though this issue is not pervasive and we can select assemblies and strains with higher quality annotations, there are likely to be discrepancies for individual genes. One solution is to manually curate the results, but for large numbers of gene comparisons or in studies that include comparisons across many species, this is not feasible. To more accurately estimate orthologous relationships, we can add a verification step to align the protein to nucleotide genomes using tblastn (Altschul et al.; 1990; Camacho et al.; 2009). We can also improve accuracy by using multiple strains in addition to reference strains for each species. We plan to include these expansions to our tools in the future.

Another important caveat to note in our study is that the transcription factors we studied play functional roles in processes other than biofilm formation. Therefore, some of our observations such as colocalization or cooperative binding could also be driven by functions outside of the biofilm regulatory network. Additionally, the biofilm regulatory network itself is tightly coupled with other processes such as the commensal-pathogen transition and morphological switching between cell types (Rodriguez et al.; 2020). Hence, it is important to consider the pleiotropic roles of these transcription factors (Kittelmann et al.; 2018) to better understand the constraints they evolve under.

It is also worth mentioning that not all interspecific binding events predicted using ChIP-Seq data and intraspecific binding site losses and gains observed using genomic data have functional consequences. Other factors such as co-binding of other regulators, chromatin accessibility and local genomic context also play roles in determining functional consequence of binding (Jiang and Mortazavi; 2018). Hence, we would need additional data, such as gene expression profiles and chromatin structure to verify our findings. A recent study conducted on a hybrid yeast strain of *Saccharomyces cerevisiae* and *S. paradoxus* showed that by combining information on motif preference and contextual binding of transcription factors, we can predict the effect of variants in binding sites on changes in expression of target genes (Krieger et al.; 2022). Thus, by including gene expression information to our motif based predictions, we can associate changes in *cis*-regulatory regions to changes in expression of downstream genes.

5.3 Future work

In light of the limitations and key findings in this study, below we present exciting future directions and developments to follow up on. We inferred biofilm regulatory networks across several *Candida* species and identified genetic variants in *C. albicans* strains spanning multiple clades and host niches. This data can be leveraged to further understand *Candida* biology and developmental processes. The tools we developed can be further refined to improve accuracy and can be expanded for additional functionality. To better understand the molecular mechanisms involved in biofilm formation, we can update the biofilm regulatory network using our workflow with improved sensitivity for binding site prediction. Additionally, we can also use metrics and methods from graph theory to identify network motifs for interspecific comparisons (Alon; 2007; Benson et al.; 2016).

We identified large structural variations and hypothesize that these are present at low frequencies. We can test this hypothesis using read depths spanning the structural variants to infer variant allele frequencies (Sudmant et al.; 2015). This will help also help us estimate the heterogeneity in the population as well as plasticity in the *C. albicans* genome. Additionally, we can also verify the structural variants using long-read genomic data when available. This will help in improving annotation of the structural variants. We also collected metadata on the strains including the geographic location, host infection status, host niche and clade. We can use this information to associate clade-specific and host niche-specific variants. Such genome-wide association studies can help identify causal relationships between variants and specific phenotypes (Uffelmann et al.; 2021).

We compared the variants and binding motifs qualitatively to the mutational landscape of the motifs. We can further add quantitative metrics to describe the mutational robustness of the motif. Specifically, we can quantify the ruggedness (Van Cleve and Weissman; 2015), robustness (Vaishnav et al.; 2022), and navigability (Aguilar-Rodríguez et al.; 2017) of the landscapes and use them as predictive measures for expectation in terms of binding site losses and gains. Alternatively, we can also compute expectation of binding site losses and gains based on motif information content, background nucleotide frequencies and genomic context (Doniger and Fay; 2007). We can also model cooperative binding to estimate expected rates of losses and gains of binding sites (Tuğrul et al.; 2015). This will allow for comparisons between the deviation of our observed, empirical losses and gains of binding sites from neutral models of binding site evolution.

An additional future development of the tools presented here include refinement of methods used to compute population genetic statistics from whole-genome sequencing data. Statistics used to measure selection can be biased due to variations in sequencing depth between genomic regions. As part of future work, we will address these biases by accounting for read depth variations between different sites and strains (Korneliussen et al.; 2013). We also plan to incorporate other statistics commonly used to measure positive selection in *cis*-regulatory regions in the future (Moon et al.; 2019).

5.4 Bibliography

Aguilar-Rodríguez, J., Payne, J. L. and Wagner, A. (2017). A thousand empirical adaptive landscapes and their navigability, *Nature ecology & evolution* 1(2): 0045.

Alon, U. (2007). Network motifs: theory and experimental approaches, *Nature Reviews Genetics* 8(6): 450–461.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool, *Journal of molecular biology* 215(3): 403–410.

Anderson, F. M., Visser, N. D., Amses, K. R., Hodgins-Davis, A., Weber, A. M., Metzner, K. M., McFadden, M. J., Mills, R. E., O'Meara, M. J., James, T. Y. et al. (2023). Candida albicans selection for human commensalism results in substantial within-host diversity without decreasing fitness for invasive disease, *Plos Biology* 21(5): e3001822.

Arnosti, D. N. and Kulkarni, M. M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?, *Journal of cellular biochemistry* 94(5): 890–898.

Benson, A. R., Gleich, D. F. and Leskovec, J. (2016). Higher-order organization of complex networks, *Science* 353(6295): 163–166.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L. (2009). BLAST+: architecture and applications, *BMC bioinformatics* 10: 1–9.

Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., Hadzic, T., Damani, F. N., Ganel, L., Consortium, G. et al. (2017). The impact of structural variation on human gene expression, *Nature genetics* 49(5): 692–699.

Doniger, S. W. and Fay, J. C. (2007). Frequent gain and loss of functional transcription factor binding sites, *PLoS computational biology* 3(5): e99.

Jiang, S. and Mortazavi, A. (2018). Integrating ChIP-seq with other functional genomics data, *Briefings in functional genomics* 17(2): 104–115.

Kittelmann, S., Buffry, A. D., Franke, F. A., Almudi, I., Yoth, M., Sabaris, G., Couso, J. P., Nunes, M. D., Frankel, N., Gómez-Skarmeta, J. L. et al. (2018). Gene regulatory network architecture in different developmental contexts influences the genetic basis of morphological evolution, *PLoS genetics* 14(5): e1007375.

Korneliussen, T. S., Moltke, I., Albrechtsen, A. and Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data, *BMC bioinformatics* 14(1): 1–14.

Krieger, G., Lupo, O., Wittkopp, P. and Barkai, N. (2022). Evolution of transcription factor binding through sequence variations and turnover of binding sites, *Genome Research* 32(6): 1099–1111.

Martin, H., Kavanagh, K. and Velasco-Torrijos, T. (2021). Targeting adhesion in fungal pathogen Candida albicans, *Future Medicinal Chemistry* 13(03): 313–334.

Moon, J. M., Capra, J. A., Abbot, P. and Rokas, A. (2019). Signatures of recent positive selection in enhancers across 41 human tissues, *G3: Genes, Genomes, Genetics* 9(8): 2761–2774.

Nather, K. and Munro, C. A. (2008). Generating cell surface diversity in Candida albicans and other fungal pathogens, *FEMS microbiology letters* 285(2): 137–145.

Rodriguez, D. L., Quail, M. M., Hernday, A. D. and Nobile, C. J. (2020). Transcriptional circuits regulating developmental processes in Candida albicans, *Frontiers in Cellular and Infection Microbiology* 10: 605711.

Staller, M. V. (2022). Transcription factors perform a 2-step search of the nucleus, *Genetics* 222(2): iyac111.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M. et al. (2015). An integrated map of structural variation in 2,504 human genomes, *Nature* 526(7571): 75–81.

Tuğrul, M., Paixao, T., Barton, N. H. and Tkačik, G. (2015). Dynamics of transcription factor binding site evolution, *PLoS genetics* 11(11): e1005639.

Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T. and Posthuma, D. (2021). Genome-wide association studies, *Nature Reviews Methods Primers* 1(1): 59.

Vaishnav, E. D., de Boer, C. G., Molinet, J., Yassour, M., Fan, L., Adiconis, X., Thompson, D. A., Levin, J. Z., Cubillos, F. A. and Regev, A. (2022). The evolution, evolvability and engineering of gene regulatory DNA, *Nature* 603(7901): 455–463.

Van Cleve, J. and Weissman, D. B. (2015). Measuring ruggedness in fitness landscapes, *Proceedings of the National Academy of Sciences* 112(24): 7345–7346.

Weingarten-Gabbay, S. and Segal, E. (2014). The grammar of transcriptional regulation, *Human genetics* 133: 701–711.