Genetic diversity of *Candida albicans* and *Nakaseomyces glabratus* across geographic scales and within-host populations

Abdul-Rahman Adamu Bukari

Department of Microbiology
University of Manitoba
Winnipeg, Manitoba, Canada

A Thesis submitted to the Faculty of Graduate and Postdoctoral Studies of the University of Manitoba in partial fulfilment of the requirements of the degree of Doctor of Philosophy

Abstract

Candida albicans and *Nakaseomyces glabratus* are major opportunistic fungal pathogens, yet both present persistent challenges in the clear delineation of clades. In *C. albicans*, phylogenomic studies have identified major clades, but often rely on unphased genomes, limited ecological and geographic sampling, and non-standardized phylogenetic methods. In *N. glabratus*, population structure remains unresolved due to ongoing discussion over whether multilocus sequence typing (MLST) or whole-genome sequencing (WGS) provides more accurate clustering. Addressing these limitations is fundamental for determining intraspecies diversity and facilitating comparative analyses. In this thesis, I employ comparative genomics from WGS of many isolates paired with computational statistics to quantify phylogenetic relationships and measure diversity. Chapter Two determines the global population structure of 1178 global *C. albicans* isolates collected from many isolation sources. I used a systematic, threshold-based clustering approach on the phased genome sequences to refine existing clade definitions and identify six novel clades. Isolates from different continents and isolation sources were not equally distributed among clades, yet these factors were confounded due to biased isolate sampling. Genomic features such as heterozygosity, mating-type locus genotype, the presence of chromosomal aneuploidy and large copy number variations, and potential for RNA interference varied widely among isolates but were generally not associated with clade. The analysis of intra-host isolates was consistent with largely clonal colonization, yet some individuals from different infectious contexts harbored isolates from multiple clades. These results reveal the global population structure of *C. albicans* and highlight the need for comprehensive, unbiased sampling to better understand its ecological and clinical diversity. In Chapter Three, I examined the population structure of 548 isolates *N. glabratus* from 12 countries. I compared clusters determined from WGS data and MLST data; I found strong concordance between the methods and proposed a new naming system to link 27 statistically supported WGS clusters with their dominant MLST sequence type. Admixture analysis identified 65 isolates with multiple ancestries: 7 are singletons, while 58 are from six different clusters. This suggests both ancient and ongoing recombination. Aneuploidy was found in 4% of the isolates, most often involving an extra copy of chromosome E, which contains *ERG11* which codes for an enzyme that is a primary target of azole antifungal drugs. These findings provide a unified framework for *N. glabratus* population classification and offer insights into its genetic diversity and gene flow within the species In Chapter Four, I move from global diversity to intra-host diversity, in the context of recurrent vulvovaginal candidiasis. I analyzed 24 rectal and vaginal isolates from three individuals with *C. albicans* infections and one with an *N. glabratus* infection. The isolates were placed into species-level phylogenies and tested for drug response and invasive growth. Genomic similarity across body sites suggested ongoing migration. Genetic variation was minimal, driven primarily by single-nucleotide changes, with little phenotypic divergence between sites. Together, these chapters provide an integrated view of fungal population structure in two important species, from global to within-host. This work proposes standardized approaches for genomic analysis across diverse contexts and highlights critical gaps in our existing isolate collection that hinders our ability to comprehensively identify the impact of geography and body site of isolation on genomic diversity.

Acknowledgements

He who does not thank the people is not thankful to Allah — Muhammad ibn Abdullah

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Aleeza Gerstein. Her unwavering support, guidance, and mentorship have been instrumental in the successful completion of this work. I am truly indebted to her for her compassion, encouragement, and for providing an environment where I was able to grow both personally and professionally. She granted me the flexibility and freedom to think independently and develop as a scientist, always offering support during challenging moments in both research and life. Words fall short in conveying how thankful I am. Dr. Gerstein not only helped me build a strong foundation in research but also guided me through the more nuanced aspects of scientific inquiry, teaching me when to push forward and when to apply the brakes. She offered invaluable advice on navigating the vast array of analyses

I was eager to pursue and helped me learn how to critically review scientific literature with a discerning eye. She also gave me the opportunity to work on a wide range of exciting projects involving organisms such as *Escherichia coli*, *Candida albicans,* and *Nakaseomyces glabratus*. While not all of these projects made it into this thesis, each of them contributed significantly to my development as a researcher. I was also fortunate to work at the Broad Institute, one of the world's most prestigious research centers, which greatly enriched my scientific training. Additionally, I improved my coding proficiency and gained hands-on experience with high-performance and cloud computing systems.

I am also profoundly grateful to the members of my PhD committee: Dr. Matthew Bakker, Dr. Colin Garroway, and Dr. Georg Hausner. Your expertise, thoughtful feedback, and constructive criticism have played a critical role in shaping and refining my research. I could not have asked for a better, more supportive committee. I would like to sincerely thank Dr. Markus Stein, Dr. Vanessa Poliquin, and all of the clinical staff and participants from THRIVE yeast for their invaluable support in facilitating the clinical strain collections from Manitoba. This thesis would not have been possible without their involvement. In particular, having Dr. Stein on board from the early stages helped shape the direction of the project in critical ways; without his early collaboration, this thesis would have taken an entirely different form. Also, I want to thank Dr Christina Cuomo from Broad institute for her guidance during my internship.

I want to thank the past and present members my amazing lab the Microstats Lab: Rebecca, Javier, Aruni, Parul, and Yana. You've all helped create a fantastic, collaborative, and supportive environment that made the lab a great place to learn and grow. I'm especially grateful to Ola for her continuous support. Our conversations, teamwork, and thoughtful discussions made my time in the lab not only intellectually stimulating but genuinely enjoyable. Thank you for always being willing to share your knowledge, offer help, and provide encouragement when I needed it most.

I'm also thankful to the Funlab members for their insightful and constructive feedback on my conference presentations. Your input challenged me to think critically and improved the way I communicate my work.

I will like to extend my thanks to the faculty, students, and staff in the Department of Microbiology. The supportive culture of the department has been essential to my PhD journey. A special thank you to our dedicated office coordinators Stephanie, Jo, and Kerry) for always being helpful, responsive, and kind. Your efforts made navigating the administrative side of things smooth and stress-free.

My PhD was supported by SEGS, GETS, UMGF, MITACs and EvoFunPath funding. I'm also grateful for the travel bursaries that enabled me to attend conferences and internships, where I had the chance to share my research with diverse scientific communities and gain broader perspectives.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AIDS	Acquired Immunodeficiency Syndrome
ANOVA	Analysis of Variance
BA	Bric Acid
BSI	Blood Stream Infection
Chr	Chromosome
CLT	Clotrimazole
CNVs	Copy Number Variations
COVID-19	Coronavirus Disease 2019
DDA	Disk Diffusion Assay
DNA	Deoxyribonucleic Acid
DST	Diploid Sequence Type
FLC	Fluconazole
FoG	Fraction of Growth
GTR+G	General Time Reversible + Gamma model
GWAS	Genome-Wide Association Study
HIV	Human Immunodeficiency Virus
LOH	Loss of Heterozygosity
MCZ	Miconazole
NCBI	National Center for Biotechnology Information
mDNA	Mitochondrial DNA
MIC	Minimum Inhibitory Concentration
ML	Maximum Likelihood
MLST	Multilocus Sequence Typing
MTL	Mating Type-Like Locus
MRE	Majority Rule Extended
NYT	Nystatin
OD	Optical Density
PBS	Phosphate Buffered Saline
RAD	Radius of the Zone of Inhibition
rDNA	Recombinant DNA
RNA	Ribonucleic Acid
RNAi	RNA Interference
RPMI	Roswell Park Memorial Institute
RVVC	Recurrent Vulvovaginal Candidiasis
SDA	Sabouraud Dextrose Agar
SNPs	Single-Nucleotide Polymorphisms
ST	Sequence Types
tRNA	Transfer RNA

UPGMA	Unweighted Pair Group Method with Arithmetic Mean
VCF	Variant Call Format
VVC	Vulvovaginal Candidiasis
WHO	World Health Organization
WGD	Whole Genome Duplication
WGS	Whole Genome Sequencing
YPD	Yeast Extract Peptone Dextrose

Contributions of authors

Chapter 1: Section 1.2.3 has been expanded and is currently in preparation for submission as a review manuscript. A.-R.A.B. conceived this study and wrote the first draft. A.-R.A.B. and A.C.G. edited the additional drafts.

Chapter 2: A.-R.A.B. and A.C.G. conceived of the study. A.-R.A.B. and A.C.G designed the study. C.H. and D.O contributed environmental isolates, C.C. hosted A.-R.A.B at the Broad Institute and contributed conceptually to the bioinformatic analyses, A.-R.A.B. harvested raw data from repositories and built the pipelines for bioinformatic analysis, A.-R.A.B. and A.C.G generated figures, conducted statistical analysis, contributed to data interpretation and wrote and edited the first and subsequent manuscript drafts.

Chapter 3: A.-R.A.B. and A.C.G. conceived and designed the study. A.-R.A.B. harvested raw data from repositories, ran the pipeline for bioinformatics, built the pipelines for data analysis, generated figures, and conducted statistical analysis. B.S. built trial pipelines and scripts for bioinformatic and data analyses as part of her fourth-year project, A.-R.A.B. and A.C.G. contributed to data interpretation. A.-R.A.B. wrote the first draft of the manuscript, A.-R.A.B. and A.C.G. edited subsequent manuscript drafts.

Chapter 4: A.-R.A.B., R.J.K.-G. and A.C.G. conceived and designed the study. V.P., R.J.K.-G. and Y.S. administered and collected clinical data. A.-R.A.B. harvested raw data from repositories and ran the pipeline for bioinformatics. A. d.G, A.S., B.M., and D.H. conducted phenotypic analyses. A.-R.A.B. and A.C.G. generated figures, conducted statistical analyses, contributed to data interpretation and wrote edited the first and subsequent manuscript drafts (generally split between genomic components– A.-R.A.B. and phenotypic components– A.C.G).

Chapter 1: Literature review

1.1 Introduction

Fungi are a diverse kingdom of eukaryotic heterotrophs that share fundamental cellular characteristics with both animals and plants, including membrane-bound organelles, a defined nucleus, and linear chromosomes. As eukaryotic organisms, fungi have molecular machinery that is distinct from prokaryotes, including bacteria and archaea. Compared to other eukaryotes, fungi are characterized by chitinous cell walls, absorptive heterotrophy, planktonic (yeast) and filamentous growth forms such as hyphae and mycelia. Unlike animals, which ingest food, fungi secrete enzymes to digest organic matter externally and absorb the resulting nutrients. They store energy as glycogen (as in animals), in contrast to the starch storage of plants. Reproduction typically involves either budding or spore formation, and many fungi exhibit a dikaryotic stage (which is distinctive to fungi compared to other eukaryotes) in their life cycle. While generally non-motile, fungi occupy a wide range of ecological niches as decomposers, symbionts, and pathogens (Feofilova, 2001).

Fungi are ubiquitous across the planet, having been detected in extreme and varied environments (Coleine et al., 2022), including the stratosphere (Wainwright et al., 2003), the hypersaline sediments of the Dead Sea (Buchalo et al., 1998; Oren and Gunde-Cimerman, 2012), deep-sea sediments (Nagahama et al., 2011), Antarctic glaciers (Freeman et al., 2009), arid deserts (Gonçalves et al., 2016), and even within the gut microbiota of flies (Blackwell, 2017). Across ecosystems, fungi perform diverse ecological roles, notably in decomposition, nutrient cycling, and the facilitation of nutrient transport. Furthermore, they engage in a variety of symbiotic relationships, such as those seen in lichens, and also include numerous species that act as pathogens affecting plants, animals, and humans.

The number of fungal species has been recently evaluated to be within the range of 2.2 and 3.8 million, yet only about 150,000 species have been described (Blackwell, 2011; Hawksworth, 2001; Hawksworth and Lücking, 2017). Currently, the kingdom Fungi is organized into nine phylum-level clades: Opisthosporidia, Chytridiomycota, Neocallimastigomycota, Blastocladiomycota, Zoopagomycota, Mucoromycota, Glomeromycota, Basidiomycota, and Ascomycota (Naranjo-Ortiz and Gabaldón, 2019). Ascomycota, the largest fungal phylum, encompasses approximately two-thirds of all described fungal species (Lutzoni et al., 2004; Schoch et al., 2009). The vast diversity of Ascomycota, combined with the relative ease of experimental manipulation, has established this phylum as a preeminent group for eukaryotic research. Several species have been utilized as model organisms, including *Candida albicans* (Kabir et al., 2012), *Saccharomyces cerevisiae* (Karathia et al., 2011; Nielsen, 2019), *Neurospora crassa* (Honda et al., 2020; Roche et al., 2014), and *Schizosaccharomyces pombe* (Hoffman et al., 2015), facilitating significant breakthroughs in understanding eukarya biology (Aramayo and Selker, 2013; Denoncourt and Downey, 2021; Dinh and Bonnefoy, 2024; Fu et al., 2008; Galagan et al., 2005; Legrand et al., 2019).

Despite the high estimate of fungal species, the vast majority of human fungal infections and associated mortalities are attributed to only a few hundred pathogenic species, which are concentrated within a limited number of taxonomic groups (Rokas, 2022). Fungal infections account for a significant global health burden, affecting more than 300 million people annually (Bongomin et al., 2017; Brown et al., 2012; Richardson, 2005). Fungal infections cause over 1.6 million deaths a year, with a mortality rate comparable to tuberculosis and higher than malaria. The mortality rate for certain invasive infections, such

as aspergillosis in some populations, surpasses 50% (Bongomin et al., 2017; Brown et al., 2012; Latgé and Chamilos, 2019; Richardson, 2005; Smith et al., 2025). However, fungal infectious disease research has historically been eclipsed by the study of bacterial and viral pathogens. This is partly due to the opportunistic nature of most pathogenic fungi; the majority primarily affect immunocompromised individuals, though their incidence is increasing due to factors such as cancer therapies, organ transplantation, HIV/AIDS, and severe viral infections like COVID-19. Many fungal virulence traits are not necessarily specialized for human infection but may reflect adaptations for survival in natural environments. In *Cryptococcus neoformans*, for example, such traits are likely pleiotropic, conferring fitness both in environmental niches and within the human host. This idea (termed "dual-use" or "accidental virulence") suggests that pathogenicity can arise incidentally from traits evolved for environmental fitness, not direct adaptation to mammalian hosts (Casadevall et al., 2003; Casadevall and Pirofski, 2007). While clinical research clarifies how these factors contribute to disease, ecological and evolutionary perspectives help explain their origins and persistence beyond the context of human infection.

In this thesis I focus on two of the most common fungal pathogens of humans, *Candida albicans* and *Nakaseomyces glabratus*. These species diverged from a common ancestor approximately 200 million years ago, and both are frequently associated with invasive infections, particularly in immunocompromised individuals. They are widely studied due to their high global prevalence. Notably, both are included on the World Health Organization's list of priority fungal pathogens (WHO, 2022).

1.2 Overview of the genus *Candida*

The genus *Candida*, initially delineated by Berkhout in 1923, circumscribes a taxonomic grouping of ascomycete yeasts characterized by asexual reproduction through budding and an absence of ascospore formation, thus distinguishing them within the order Saccharomycetales (Berkhout, 1923). Historically, the genus *Candida* encompassed a heterogeneous assemblage of morphologically similar yeasts, isolated from a diverse spectrum of ecological niches, including soil, plant matter, and the mucosal surfaces of animals. Within this broad classification, *C. albicans* was a common species recognized by the late nineteenth century as an etiological agent of both mucosal and invasive human infections (Knoke and Bernhardt, 2006; Roberts, 1988).

Throughout the twentieth century, the use of predominantly morphological criteria in the taxonomic delineation of *Candida* led to a polyphyletic genus, comprising over 200 species with often limited phylogenetic congruence. The advent of molecular phylogenetic methodologies in the latter decades of the century catalyzed a substantial taxonomic revision (Daniel et al., 2014; Guzmán et al., 2013; Kurtzman and Robnett, 1994). Genomic analyses revealed that numerous species previously subsumed within *Candida* belonged to discrete evolutionary lineages (Kurtzman and Robnett, 2003), necessitating their reclassification into more phylogenetically coherent genera, including *Clavispora* (Rodrigues de Miranda, 1979), *Meyerozyma* (Kurtzman and Suzuki, 2010), and *Debaryomyces* (Kurtzman, 2011). A total of 21 species previously classified under the genus *Candida* were renamed and reassigned to other genera as part of a major taxonomic revision proposed between 2018 and 2019 (Borman and Johnson, 2023). Among the species that retained the *Candida* genus name are some of the most frequently isolated yeasts in clinical settings, such as *C. albicans*, *Candida*

auris, *Candida dublinensis*, and *Candida tropicalis*, which belong to the CTG clade (Santos et al., 2011). Yeasts in this clade exhibit a unique genetic trait: the CTG codon is predominantly translated as serine rather than leucine. By contrast, *Nakaceomyces glabratus* (formerly *Candida glabrata*), the second most common human fungal pathogen in many jurisdictions, has been recently moved to a different genus (Borman and Johnson, 2021). The contemporary application of the genus name *Candida* thus reflects a historically pragmatic, yet increasingly phylogenetically untenable, classification system. This taxonomic realignment has generated tension within the clinical community, where stability in naming is often prioritized for diagnostic consistency and communication, sometimes in conflict with phylogenetic accuracy (Denning, 2024; Kidd et al., 2023). This system is presently undergoing a gradual but discernible transition towards a taxonomic framework predicated on evolutionary coherence and substantiated by robust genomic datasets.

1.3 *Candida albicans*

C. albicans exhibits morphological plasticity and is capable of transitioning between budding yeast cells, pseudohyphae, and true hyphae depending on the environment (reviewed in Sudbery et al., 2004). Yeast cells are typically oval to ellipsoid and reproduce via multilateral budding. Pseudohyphae are characterized by elongated cells with constrictions at septal junctions, whereas true hyphae are continuous, septate filaments with parallel walls. Under nutrient-limiting conditions, *C. albicans* may also produce chlamydospores (Staib and Morschhäuser, 2007), thick-walled, spherical structures commonly observed at the termini of pseudohyphae. Compared to other yeasts, a distinctive feature of *C. albicans* is its capacity for white-opaque switching (Lockhart et al., 2002; Lohse

and Johnson, 2009), a heritable and reversible transition between the two distinct cell types. White-phase cells are round and form smooth, domed colonies, while opaque-phase cells are elongated and produce flatter, duller colonies. This switching phenomenon is epigenetically regulated (Zordan et al., 2006) and associated with mating competence and acclimation, contributing to survival and propagation.

The *C. albicans* diploid genome consists of approximately 14.5 Mb of nuclear DNA organized across eight chromosomes, encoding approximately 6,000 protein-coding genes (Muzzey et al., 2013). The *C. albicans* genome includes several expanded gene families involved in adhesion, proteolysis, and other virulence-associated functions, relative to *C. tropicalis* (Butler, 2014). *C. albicans* predominantly reproduces by mitotic (i.e., clonal) division. Unlike *S. cerevisiae*, *C. albicans* is thought to lack a conventional sexual cycle and instead undergoes a parasexual cycle (Bennett, 2015), in which diploid cells of opposite mating types fuse, forming a tetraploid cell that is subsequently diploidized by concerted chromosomal loss. Like other primarily asexual fungal microbes, *C. albicans* is tolerant of ploidy variation and aneuploidy (Bennett, 2015; Berman and Hadany, 2012). Although most species-wide genetic variation in *C. albicans* likely arises from mutations during mitotic growth, the parasexual cycle and mitotic recombination can also generate genetic diversity, particularly under stress or selective pressure, where such mechanisms may provide fitness advantages over euploid parent strains (Yang et al., 2021).

The *C. albicans* genome exhibits chromosome structural polymorphism in chromosome length (primarily through the contraction or expansion of repeat regions, (Todd et al., 2019), reciprocal translocations (Chibana et al., 2000), heterozygosities (Marton et al., 2021), and copy number variation (Hirakawa et al., 2015). These genomic changes have

been linked to changes in phenotypic traits, including antifungal resistance, altered morphogenesis, and host adaptation. Compared to other yeasts such as *C. tropicalis*, *C. albicans* maintains a relatively high level of genome-wide heterozygosity, though this varies among isolates (O'Brien et al., 2021). Loss of heterozygosity (LOH), which converts heterozygous loci to homozygous ones, is a major genomic mechanism that can expose recessive alleles (Feri et al., 2016; Gerstein et al., 2014). LOH occurs through processes such as mitotic recombination, break-induced replication, and chromosome loss (Ene et al., 2018; Feri et al., 2016). The specifics of the biotic environment have been shown to influence the rate of LOH and the position of LOH tracts (Diogo et al., 2009; Ene et al., 2018; Forche et al., 2011; Rosenberg, 2011). Large-scale LOH events can affect entire chromosomal arms or whole chromosomes, yet short-tract LOH events (<5 kb) are more common during microevolution. Importantly, LOH can drive clinically relevant phenotypes, including antifungal resistance (Coste et al., 2006; Dunkel et al., 2008; Ford et al., 2015; White, 1997) and host-specific fitness advantages (Liang et al., 2019), making it a key mechanism of genome evolution and adaptation in *C. albicans*.

In addition to its nuclear genome, *C. albicans* has a circular mitochondrial genome of approximately 41 kilobases. It encodes key components of the respiratory chain, along with ribosomal RNAs and tRNAs necessary for mitochondrial protein synthesis (Kolondra et al., 2015). Mitochondrial function is closely tied to several aspects of pathogenicity, including morphogenesis, biofilm formation, and stress response. As such, the mitochondrial genome, though compact, plays a critical role in the survival of *C. albicans*.

1.3.1 Ecological niches of *Candida albicans*

C. albicans is commonly found as a commensal yeast that colonizes mucosal surfaces in humans and other warm-blooded animals, the oral cavity, gastrointestinal tract, and vaginal mucosa. Its presence under certain conditions can be beneficial to the host. For instance, in the intestine, *C. albicans* colonization promotes the expansion of fungal-specific Th17 $CD4^+$ T cells (Kashem et al., 2015; Shao et al., 2019) and enhances IL-17 responsiveness in circulating neutrophils, contributing to protection against systemic fungal infections. Furthermore, commensal fungi can compensate for the absence of certain intestinal bacteria by protecting mucosal tissues from injury and modulating the activation state of peripheral immune cells, suggesting a broader role in maintaining immune homeostasis (Jiang et al., 2017).

The ability of *C. albicans* to colonize diverse anatomical sites reflects its ability to acclimate to various environmental conditions. This includes variation in pH, ranging from approximately 3 in the stomach to 8 in the small intestine, (Barbosa et al., 2020; Davis, 2003), nutrient levels from simple sugars to complex carbohydrates and proteins, oxygen availability from aerobic conditions in the oral cavity to anaerobic in the colon (Burgain et al., 2020; Zheng et al., 2015), and host immune responses (Stappers and Brown, 2017). In the oral cavity, *C. albicans* is frequently detected on the tongue and cheeks, in the saliva, and within a complex microbial biofilm of dental plaque (Patel, 2022). Within the gastrointestinal tract, *C. albicans* occupies regions such as the esophagus, stomach, and intestines, often forming biofilms, which can influence its interaction with the host and other microbes.

C. albicans has traditionally been considered an obligate human commensal, with rare environmental isolates due to human contamination. However, the studies that have

purposefully sought to isolate *C. albicans* from the environment repeatedly isolate it from soil, water, and plant material (Bensasson et al., 2019; Fotedar et al., 2022; Opulente et al., 2019; Sautour et al., 2021; Stone et al., 2012; Yamaguchi et al., 2007), suggesting that *C. albicans* is likely capable of stably persisting outside a human host. Survival in ecological niches could imply novel exposure routes for human infection (as has been shown in for other human pathogens, van Rhijn and Rhodes, 2025), especially since environmental isolates have previously been shown to be genetically similar to clinical isolates (Bensasson et al., 2019). However, the environmental conditions and biological interactions that support *C. albicans* growth and persistence outside the host remain poorly characterized and warrant further investigation.

1.3.2 *Candida albicans* as a pathogen

C. albicans is a common human commensal that, under certain changes in the host's internal environment or external conditions, can transition to a pathogenic state and cause infections at various body sites. This capacity to shift between commensalism and pathogenicity is key to understanding its role in human disease, especially in immunocompromised individuals, where such transitions are more likely to result in clinically significant infections. The shift toward a pathogenic lifestyle is frequently triggered by external factors such as the widespread administration of broad-spectrum antibiotics (Seelig, 1966; Xu et al., 2008), fluctuations in hormonal balance, physical disruption of mucosal barriers, or systemic immunosuppression. These conditions can create ecological niches that favor fungal overgrowth and subsequent tissue invasion.

Several virulence factors have been identified that delineate the transition to pathogenicity. Morphological transitions facilitate tissue invasion, enable a circumvention of the host's immune surveillance mechanisms, and underpin the development of robust biofilms (Noble et al., 2017; Thompson et al., 2011). The filamentous hyphal cells, acting as invasive structures, possess the capability to actively penetrate epithelial barriers through the secretion of potent hydrolytic enzymes (Schaller et al., 2005), notably secreted aspartyl proteases and phospholipases. Concurrently, surface adhesins, such as ALS3 (Hoyer and Cota, 2016) and HWP1 (Sundstrom et al., 2002), function as molecular anchors, mediating firm attachment to host cells and diverse surfaces.

Biofilm formation also significantly enhances the pathogenic potential of *C. albicans* (Cavalheiro and Teixeira, 2018; Soll and Daniels, 2016). Biofilms form on indwelling medical devices and mucosal surfaces, and resistance to conventional antifungal agents and the host's innate immune responses, thereby posing considerable challenges in clinical management. *C. albicans* also employs sophisticated mechanisms to evade detection and elimination by the host's immune system. One such strategy involves the strategic concealment of immunostimulatory β-glucans, key structural components of its cell wall, beneath an outer layer enriched in mannan polysaccharides (Chen et al., 2022; Hameed et al., 2021). This molecular masking effectively impedes recognition by host pattern recognition receptors. Moreover, *C. albicans* exhibits the capacity to escape the destructive processes of phagocytic cells and actively modulate the host immune response (da Silva Dantas et al., 2016; Oliver et al., 2019; Zhou et al., 2021), often skewing it away from the Th1-mediated responses that are critical for effective antifungal defense.

1.3.3 Genetic diversity of *Candida albicans*

1.3.3.1 Typing *Candida albicans*

As with other medically important fungi such as *Cryptococcus neoformans*, early attempts to characterize within-species variation in *Candida albicans* relied on serotyping, due to the limited availability of sequence-based methods. By the late 20th century, three serotyping techniques had been developed, utilizing antisera HSN1 (yielding serotypes A and B) and HSN2 (Hasenclever and Mitchell, 1961a, 1961b), the Iatron *Candida* Check factor 6 (IF6, Poulain et al., 1985), and agglutination with monoclonal antibody H9 (Brawner and Cutler, 1989). While serotyping was extensively utilized (Brawner and Cutler, 1989; Odds et al., 1989; Stiller et al., 1982; Whelan et al., 1990), these methods grouped *C. albicans* into only a few categories, offering limited epidemiological resolution, and lacked consistent correlation with one another (Brawner, 1991). Even more concerning was the finding that antigen expression could vary with the growth phase, and that serotype B cells could produce serotype A antigens (Poulain et al., 1985), further undermining the validity of serotyping as a robust classification method. Recognizing the limitations of serotyping, multiple biotyping methods were developed including morphotyping (Hunter et al., 1989; Phongpaichit et al., 1987; Quindós et al., 1992), resistotyping (Hunter and Fraser, 1987; McCreight et al., 1985), killer yeast typing (Polonelli et al., 1983, 1985), enzyme typing (Román and Linares Sicilia, 1983; Williamson et al., 1986, 1987), sugar assimilation typing (Buesching et al., 1979; Fricker-Hidalgo et al., 1996), drug susceptibility typing (Quindós et al., 1996), isoenzyme biotyping (Lehmann et al., 1989) and a biotyping system based on assessment of growth patterns of yeasts on 10 agar test media (Odds and Abbott, 1980).

When the complex biotyping system was applied to oral and vaginal isolates, it was able to identify 45 types (Odds, 1997). This method was used in several epidemiological studies (Lipperheide et al., 1996; Poirier et al., 1990; Xu and Samaranayake, 1995). However, a later study across five labs revealed poor interlaboratory consistency (Odds and Abbott, 1983). As a result, the method was deemed suitable for research but unreliable for clinical diagnostics. Biotyping methods gradually became less common due to, among other reasons, the emergence of DNA fingerprinting (reviewed in (Soll, 2000).

The age of DNA fingerprinting started with non-sequencing approaches (i.e., Restriction Fragment Length Polymorphism, rDNA and mDNA probes, repetitive and complex DNA probes, Random Amplified Polymorphic DNA). DNA fingerprinting probe *Ca3* received extensive usage (Anderson et al., 1993; Marco et al., 1999; Pujol et al., 1999). This was later replaced with sequencing approaches such as, SNP array analysis, polymorphic microsatellite typing and multilocus sequence typing (MLST, also referred to as a diploid sequence type, DST, in *C. albicans*). Sequence-based typing systems were shown to give concordant results in several studies, and the MLST approach continues to serve as a common approach.

Early multilocus sequence typing (MLST) efforts for *Candida albicans* initially led to the development by different research groups of separate sets of single-nucleotide polymorphisms (SNPs, Bougnoux et al., 2002; Tavanti et al., 2003). To improve consistency and comparability across studies, two groups collaborated to establish a unified and discriminatory seven-gene set (*AAT1a*, *ACC1*, *ADP1*, *MPIb*, *SYA1*, *VPS13*, and *ZWF1b*), which still forms the foundation for most MLST-based population studies of *C. albicans* (Bougnoux et al., 2003). The MLST typing scheme includes genes on six of the eight chromosomes, with

chromosomes 3 and 5 unrepresented. A proposed more extensive MLST approach involved sequencing 24 genes, strategically selecting one gene located at the center and one at each arm of all eight chromosomes (Odds, 2010). However, with the rise of relatively cost-effective whole-genome sequencing (WGS) at approximately the same time, this expanded scheme was never widely adopted by the community, and the original seven-gene scheme remains the standard.

1.3.3.2 *Candida albicans* population structure and clades

The current *C. albicans* MLST database includes over 5,000 isolates, comprising 4339 DSTs. To reduce this diversity into phylogenetically meaningful clusters, researchers applied a pairwise (P) distance threshold of 0.04 to cluster isolates together, as proposed by Odds et al., (2007). P distance refers to the proportion of nucleotide differences between two sequences across the concatenated MLST loci. The choice of a 0.04 cutoff (i.e., isolates that differ by less than 4% of aligned nucleotide positions are grouped into the same clade) was somewhat arbitrary. This decision was justified on the basis that it separated clusters containing isolates known to belong to clades 2 and 4 (also referred to in earlier manuscripts as clade SA), which had previously been identified using DNA fingerprinting with the moderately repetitive probe *Ca3* (Soll and Pujol, 2003). Using this cutoff, a UPGMA (unweighted pair group method with arithmetic mean) phylogenetic analysis of 1,391 *C. albicans* isolates defined 17 clades (Odds et al., 2007). This clade-based classification became a widely adopted framework for studying the population structure, evolution, and epidemiology of *C. albicans.* The majority of isolates were concentrated in five dominant clades (clades 1 through 4 and clade 11). Several clades showed evidence of geographic

enrichment, though none were geographically exclusive. Clade 2 was largely composed of isolates from the UK, while clades 1 and 3 were enriched for North American isolates, and clades 14–17 were dominated by strains from the Far East. Clade 4, previously described as "African," ("SA" stands for South Africa) contained only 14.3% African isolates (fewer than those from the UK; 37.4%) yet still showed modest African enrichment relative to other major clades. However, the isolate panel used in the study was heavily skewed toward Europe and North America, so caution is required for an interpretation of geographic enrichment. Of the 1,391 isolates, nearly 620 came from the UK and France alone (226 from Scotland, 271 from England/Wales and 123 from France), compared to just 72 from Africa, 25 from Australasia and 15 from the Middle East. This imbalance likely influenced clade composition and limits broader conclusions about the global population structure of *C. albicans*. Isolates that did not cluster with others within the 0.04 threshold and clusters containing fewer than 10 isolates were labelled singletons.

More recent studies have employed whole-genome short-read sequencing (WGS) to build on the MLST-based framework and refine the phylogenetic classification of *C. albicans*. Using WGS data from 182 isolates (including five from the initial MLST study), Ropars et al. (2018) reconstructed a phylogeny using a maximum likelihood approach. These isolates were also predominantly from Europe (113), with additional representation from Africa (32), Asia (23), South America (5), and North America (1). They were sampled from a variety of anatomical sites, including vaginal (61), urogenital (29), oral (26), circulatory (26), gastrointestinal (8), skin (3), and environmental (21) sources. Ropars et al. (Ropars et al., 2018) recovered 12 of the previously defined MLST clades (Clades 1-4, 8-13, 16, 18). They additionally introduced a new naming scheme for new clades that did not have numbers,

assigning alphabetical labels to five clusters (clades A–E) that contained fewer than 10 but more than 2 isolates. While this expanded system allowed for a more inclusive classification of previously ungrouped lineages, it was not explicitly stated whether the original 0.04 P distance threshold was consistently applied to define the new alphabetical clades. and in some cases, application of this threshold might not have supported the delineation of certain clusters (Figure 1.1), suggesting some flexibility in the updated classification criteria.

Ropars et al. (2018) also assigned ten isolates previously classified as singletons into new or existing or new clades (two into previously numbered clades and the remainder into alphabetical ones). A small number of isolates also changed clade assignments: single isolates from clade 1 were reassigned to clade 3 and to clade 4; three isolates from clade 3 moved to clade A, and an isolate each from clades 4, 7, 8 and 9 were reclassified as singletons. These changes likely reflect improved phylogenetic resolution afforded by WGS, although they also highlight the dynamic nature of clade assignments when data from more isolates become available.

In the same study, Ropars et al. (2018) demonstrated ongoing gene flow and parasexual recombination across specific clades; a finding that overturned the long-held view of strictly clonal reproduction in this species. The updated naming convention introduced by Ropars et al. (2018) has also been adopted in other WGS-based studies (174 citations as of July 2025) that examine gene flow in different contexts. For example, Bensasson et al. (2019) used the system to classify three *C. albicans* isolates from oak tree bark, while Szarvas et al. (2021) applied it to clinical isolates collected in Denmark. Overall, this work fundamentally reshaped our understanding of *C. albicans* by highlighting the dual importance of clonality and genetic exchange in shaping population structure and facilitating adaptation.

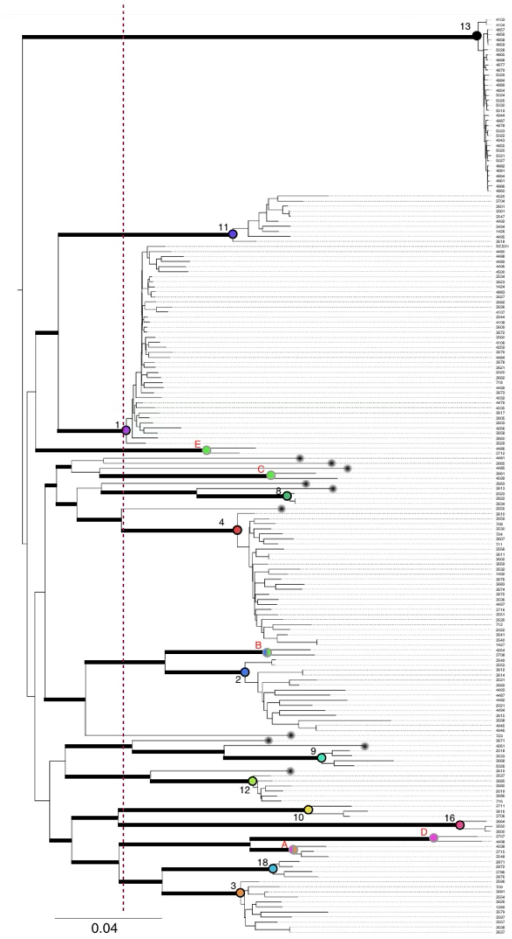


Figure 1.1: WGS phylogenetic tree showing the population structure of *C. albicans*. This figure was adapted with modification from Ropars et al. (2018). The red vertical dashed line highlights that, while Clade 1 retains a well-defined structure, the proximity of certain clades (e.g. clades 2 and B, A and D, and 3 and 18) raises the possibility that they may not be sufficiently distinct to warrant separate designations.

Recently, Gong et al. (2023) expanded on the Ropars et al. phylogeny. They incorporated 370 additional isolates from China into the Ropars dataset. Most of the isolates (n=238) clustered among existing clades. In addition to the 17 Ropars et al. (2018) clades (twelve numbered and five lettered clades), they proposed adding 21 new clades, designated as Group 1 through Group 21 (Gong et al., 2023) for a total of 38 clades. The specific criteria used to define these new clades were not clearly stated. In addition to proposing new clades, Gong *et al*. suggested that a subgroup of isolates within Clade 1 be designated as Clade 1-R,

based on shared patterns of LOH at the terminal ends of chromosomes 2 and 3. This subgroup was further subdivided into Clade 1-R-α, characterized by an additional LOH event on the terminal end of chromosome R, and Clade 1-R-β, defined by the presence of a Y132H substitution in the Erg11p protein (Gong et al., 2023). This naming scheme has not been widely adopted by the research community, yet it serves as an important reminder that within-species phylogenies are highly dependent on the isolates that are included. Taken together, while WGS has greatly improved resolution in studying *C. albicans* population structure, clade delineation practices have varied across studies. The absence of a standardized, consistently applied threshold suggests that current classifications may reflect methodological or interpretive choices as much as underlying phylogenetic relationships.

A clear and consistent clade delineation in *C. albicans* is essential for accurately interpreting its population structure, tracking the distribution of specific lineages, and drawing meaningful comparisons across studies. Inconsistent or subjective clustering criteria can lead to confusion about the evolutionary relationships between isolates and may obscure patterns related to geographic distribution. Standardizing the criteria for clade assignment, such as consistently applying a defined genetic distance threshold, would enhance the reproducibility and comparability of findings, ultimately strengthening both basic research and potential clinical surveillance efforts.

1.4 *Nakaseomyces glabratus*

Nakaseomyces is a genus of fungi within the family *Saccharomycetaceae* of the phylum Ascomycota (Kurtzman, 2003). The genus comprises six closely related species, categorized into three environmental species: *Nakaseomyces castellii* (syn. *Candida castellii*), *N.*

delphensis (syn. *Kluyveromyces delphensis*), and *N. bacillisporus* (syn. *K. bacillisporus*); and three pathogenic species: *N. bracarensis* (syn. *C. bracarensis*), *N. nivariensis* (syn. *C. nivariensis*), and *N. glabratus* (syn. *Candida glabrata*). These close genetic relationships between environmental and pathogenic species provide an opportunity for comparative studies. The most well-studied species within this genus is *N. glabratus.*

Morphologically, *N. glabratus* is a small, oval-shaped yeast that reproduces asexually through multilateral budding and grows almost exclusively in the yeast form (Kaur et al., 2005; Muller et al., 2008). On solid media, it typically forms smooth, creamy colonies, though *N. glabratus* colonies can also be light brown, dark brown and very dark brown in the presence of copper sulphate or phloxine B (Lachke et al., 2002). In clinical settings, *N. glabratus* is commonly identified using culture-based techniques, such as CHROMagar *Candida*, where it produces colonies ranging in color from mauve to pink or purple (in contrast to the green *C. albicans* colonies), and through microscopic examination, which reveals small yeast cells (approximately 1 - 4 μm) without hyphal elements.

N. glabratus is phylogenetically more closely related to *Saccharomyces cerevisiae* than to *C. albicans*, reflecting a deep evolutionary divergence between the two pathogenic yeasts. Notably, the common ancestor of *N. glabratus* and *S. cerevisiae* (which existed about 100-200 million years ago (Gabaldón et al., 2013)), underwent an ancestral whole-genome duplication (WGD) event which is absent in the lineage leading to *C. albicans* (Dujon et al., 2004). The WGD has profoundly shaped the genome architecture and regulatory complexity of *N. glabratus*, contributing to its biological and pathogenic traits (Bolotin-Fukuhara and Fairhead, 2016; Gabaldón et al., 2013). For example, the WGD resulted in expanded gene

families involved in stress response, adhesion, and drug resistance, and enabled functional divergence between some gene duplicates.

N. glabratus is haploid, with an approximately 12.3 Mb genome distributed across 13 chromosomes (Dujon et al., 2004). In contrast to the relatively streamlined genome architecture of *C. albicans*, *N. glabratus* features longer intergenic regions (Gabaldón et al., 2013), a trait associated with more complex gene regulation, transcriptional control, and chromosomal stability. Compared to *C. albicans, N. glabratus* has a smaller genome size (~12.3 Mb vs. ~14.5 Mb in *C. albicans*) and smaller protein coding genes (5272 vs. 6218; taken from http://www.candidagenome.org). Other comparable features include basic genomic metrics such as GC content (~38.8% vs. ~33.53% in *C. albicans*) and average gene length (~1479 vs. ~1,439 bp in *C. albicans*; Braun et al., 2005; Dujon et al., 2004) which may reflect structural similarities between the genomes, though not necessarily conservation at the gene level. Indeed, comparative analyses indicate that only ~30% of *N. glabratus* genes do not have identifiable orthologs in *C. albicans* (similarly, ~36% *C. albicans* genes have no *N. glabratus* orthologues; Gabaldón and Carreté, 2016), underscoring substantial divergence in gene content despite surface-level similarities. Like *C. albicans*, *N. glabratus* predominantly reproduces asexually with a predominantly clonal lifestyle and diploid and polyploid isolates have been found exhibiting variation in morphologies and virulence (Zheng et al., 2022).

The mitochondrial genome of *N. glabratus* is a small, circular DNA molecule measuring approximately 20 kilobases. It encodes a total of eleven open reading frames (ORFs), which include genes responsible for key components of mitochondrial energy metabolism. Specifically, it contains genes for three cytochrome c oxidase subunits (*Cox1*,

Cox2, *Cox1*), apocytochrome b (*Cob*), and three ATP synthase subunits (*Atp6*, *Atp8*, *Atp6*). In addition to these protein-coding sequences, the genome also includes 23 transfer RNA (tRNA) genes, two ribosomal RNA (rRNA) genes, and one non-coding RNA, indicating a compact yet functionally sufficient genetic structure for mitochondrial function (Koszul et al., 2003; Lew-Smith et al., 2025).

1.4.1 *Nakaseomyces glabratus* as a pathogen

N. glabratus has been found in various ecological habitats, including fermented coffee beans (de Melo Pereira et al., 2014), droppings of yellow-legged gulls (Al-Yasiri et al., 2016) and soil (Opulente et al., 2019). *N. glabratus* is also an opportunistic human pathogen, and is the most common species of non-albicans clinical yeasts in many places in the world (Katsipoulaki et al., 2024; Pfaller et al., 2011). Infections caused by *N. glabratus* predominantly affect immunocompromised individuals, including the elderly, diabetics, , and recipients of solid organ transplants (Barchiesi et al., 2017; Rodrigues et al., 2019). Phylogenetic and genomic studies suggest that the pathogenic potential of *N. glabratus* is likely due to pre-adaptive traits from its environmental ancestors, rather than a direct evolutionary path from human commensal organisms (Gabaldón et al., 2013; Gabaldón and Carreté, 2016). Environmental reservoirs, such as gulls, are believed to contribute to the global spread of *N. glabratus*, potentially facilitating indirect transmission to humans through contaminated ecosystems (Al-Yasiri et al., 2016).

N. glabratus exhibits a diverse array of virulence factors that enable it to persist and thrive in both commensal and pathogenic states. One of the hallmark traits of *N. glabratus* is its strong adherence to host surfaces, particularly under stress conditions or during biofilm formation, mediated by a family of epithelial adhesins (Cormack et al., 1999; Timmermans

et al., 2018) and several polymorphisms in the Sir3 (silent information regulator) protein observed in hyper-adherent isolates (Martínez-Jiménez et al., 2013). In addition to Epa-mediated adhesion, *N. glabratus* forms biofilms on medical devices and epithelial linings, which not only protect the yeast cells from host immune responses and also significantly enhance resistance to antifungal agents (Iraqui et al., 2005).

Several biological and phenotypic traits distinguish *C. albicans* and *N. glabratus* (Table 1.1). While both are major opportunistic pathogens, they differ in stress tolerance, drug resistance, and host interactions (Katsipoulaki et al., 2024). Unlike *C. albicans*, *N. glabratus* does not undergo true hyphal transformation, but it relies on other traits that promote survival and persistence in the host (Brockert et al., 2003; Lachke et al., 2002). These include robust stress response systems, efficient nutrient acquisition strategies (Kaur et al., 2005; Sprenger et al., 2020), and intrinsic resistance to certain antifungal agents, particularly azoles, mediated by the upregulation of efflux pumps and mutations in the ergosterol biosynthesis pathway. *N. glabratus* also exhibits notable tolerance to oxidative and osmotic stress (Cuéllar-Cruz et al., 2008; Roetzer et al., 2011), and it can survive within macrophages, avoiding clearance by host immune cells (Roetzer et al., 2010). Together, these features reflect a form of phenotypic versatility that supports virulence and persistence, particularly in immunocompromised individuals, although it is not necessarily greater than that of *C. albicans* but rather operates through different mechanisms.

1.4.2 Genetic diversity of *N. glabratus*

Genetic diversity among *N. glabratus* isolates has been studied by various molecular techniques, including polymorphic locus sequence typing, pulsed field gel electrophoresis

(Bennett et al., 2004; Lin et al., 2007), amplified fragment length polymorphism analysis (Paluchowska et al., 2014), multilocus microsatellite typing (Abbes et al., 2012) and multilocus sequence typing (MLST, Dodgson et al., 2003; Lott et al., 2010). The most common among these techniques is a six-gene MLST scheme, which initially was used to identify five clades (Dodgson et al., 2003), three of which showed significant geographical bias. Later studies that included more strains increased the number of clades to seven (Dodgson et al., 2005). MLST clade designation was fairly consistent with clades defined through microsatellite typing (Bordallo-Cardona et al., 2019; Gabaldón et al., 2020).

Table 1.1: Comparison of morphological and pathogenesis traits of *N. glabratus* and *C. albicans*

Feature	***N. glabratus***	***C. albicans***
Ploidy	Haploid	Diploid
Cellular morphology	Mainly yeast	Yeast, pseudohyphae and hyphae
Cell size	1–4 μm	4–6 μm
Phylogeny	Non-CTG clade (WGD clade)	CTG clade
Phenotypic switching	Present (Core switching system such as irregular wrinkle switching)	Present (White-opaque switching)
Carbon assimilation	Glucose and trehalose	Glucose, trehalose, maltose and galactose
Auxotrophy	Niacin, thiamine, pyridoxine	None
Crabtree effect (preferential fermentation of glucose to ethanol even when oxygen is available and sufficient for respiration)	Positive	Negative
Mitochondrial function	Petite positive	Petite negative
Mating loci	Three (*MTL1*, *MTL2*, *MTL3*)	One (MTL)
Mating loci chromosomal location	*MTL1* and *MTL3* on ChrB; *MTL2* on ChrE	MTL locus on Chr5
HO endonuclease site	Present within α1 gene	Absent
Haem receptor	Absent	Present
Haemoglobin and transferrin utilization	Absent	Present
Innate azole resistance	Present	Absent
Secretory aspartyl proteases	Absent	Present
Lifestyle	Pathogenic and potentially commensal	Common commensal and opportunistic pathogen
Major adhesins	Lectins (Epa)	Lectins (Als and Hwp)
Biofilm	Present	Present
Invasion	Not known	Induced endocytosis and active penetration
Damage to host cells	No significant damage	Substantial damage

The first published maximum-likelihood phylogeny based on WGS data from 34 isolates also revealed seven genetically distinct clades (Carreté et al., 2018). While these broadly aligned with clades defined by MLST, the topologies did not fully overlap, demonstrating the improved resolution of WGS for inferring isolate relatedness. In contrast to earlier MLST-based studies, which showed some degree of geographical clustering, the WGS phylogeny lacked strong geographic structure. Notably, the *N. glabratus* phylogeny revealed deeper intra-clade divergence than what is typically observed in *C. albicans*, both the initial WGS tree (Carreté et al., 2018) and a more recent phylogeny based on 151 isolates (Helmstetter et al., 2022) provided evidence of population admixture, suggesting that recombination, potentially from an unrecognized sexual cycle, may be shaping the *N. glabratus* population structure.

Critically, isolates in all published studies have been primarily drawn from Australia, the United Kingdom, and the United States, leaving large geographic regions underrepresented. Understanding how isolates from such understudied regions relate to global clades is important for monitoring *N. glabratus* evolution and transmission. Including regional populations improves the resolution of global phylogenies and may reveal novel clades, recombination patterns, or transmission dynamics that are otherwise missed. This approach helps ensure a more complete understanding of *N. glabratus* population biology, information that is essential for guiding public health surveillance and clinical decision-making.

Despite its clinical importance, relatively few studies have explored intrapopulation genetic variation in *N. glabratus*, particularly within individual infections. Recent genomic analyses have begun to reveal that bloodstream infections may involve clonal populations

with underlying genetic and phenotypic diversity. Studies examining serial and co-isolated strains from single patients have identified variation in genes associated with drug resistance, cell wall structure, and virulence, including non-synonymous mutations and chromosome copy number variation that arise during infection (Badrane et al., 2023; Carreté et al., 2019). These findings indicate that *N. glabratus* can undergo genetic diversification within the host, potentially contributing to treatment failure and persistence. Nevertheless, data on within-host evolution remain limited, and broader investigations are needed to fully understand the extent and clinical relevance of this variation.

1.5 Recurrent vulvovaginal candidiasis as a case study

1.5.1 Vulvovaginal candidiasis

Vulvovaginal candidiasis (VVC) is a common mucosal fungal infection affecting approximately 75% of women at some point during their reproductive years. The condition is typically characterized by symptoms such as vulvovaginal pruritus, erythema, discharge, and discomfort. While *C. albicans* is the principal etiological agent, non-albicans species such as *N. glabratus* are increasingly identified. Most cases of VVC are sporadic and respond well to antifungal treatment; however, approximately 8% of afflicted women experience recurrent symptomatic infections.

1.5.2 Recurrent vulvovaginal candidiasis (RVVC)

Recurrent vulvovaginal candidiasis (RVVC) is defined clinically as the occurrence of 3 or more symptomatic episodes of vulvovaginal candidiasis within 12 months (Farr et al., 2021; Sherrard et al., 2011; van Schalkwyk et al., 2015). RVVC is a significant gynecological concern, contributing to reduced quality of life and increased healthcare burden (Ehrström

et al., 2007; Giraldo et al., 2012; Nyirjesy et al., 2006). Like VVC, *C. albicans* is the predominant species associated with RVVC (Sobel, 2016), with an increasing incidence of *N. glabratus* and other non-albicans species (Sobel, 2007). Unlike *C. albicans*, which elicits strong inflammatory responses due to its hyphal morphogenesis and proteolytic activity, *N. glabratus* infections tend to be less symptomatic (Brunke and Hube, 2013) yet more persistent and difficult to treat due to their intrinsic azole resistance.

1.5.3 Genetic diversity in RVVC

Studies investigating genetic diversity in R/VVC (i.e., collectively VVC and RVVC) have primarily relied on multilocus sequence typing (MLST, Song et al., 2022; Tian et al., 2021; Zhu et al., 2022). These studies generally report that the same or similar *C. albicans* genotypes persist across multiple symptomatic episodes, consistent with a model of relapse. Occasionally, strain turnover (replacement of one genotype by another) has been observed, indicating possible reinfection (Schröppel et al., 1994). However, most of this work has been based on the analysis of only one or a few isolates per patient per time point. This limited sampling, as well as estimates from a handful of markers, makes it difficult to determine whether apparent reinfections are truly due to novel strains or represent standing genetic variation within a stable population that goes undetected.

WGS has rarely been applied to characterize genetic diversity in R/VVC: only one study sequenced four vaginal isolates (three *C. albicans* and one *N. glabratus*) from four women at a single time point (Bradford et al., 2017), and none have compared multiple isolates collected from the same symptomatic episode or over time. Consequently, the extent and structure of standing genetic variation in RVVC remain unknown.

1.6 Thesis research questions and hypotheses

Phylogenetic analyses provide a critical framework for interpreting genetic diversity in *C. albicans* and *N. glabratus* by clarifying evolutionary relationships among isolates and enabling the identification of well-supported clades, including emerging or previously unrecognized lineages. By placing strains within an evolutionary context, phylogenies facilitate the linkage of genetic variation to biologically and clinically relevant traits, such as differences in virulence, antifungal resistance, and patterns of within-host diversification. Moreover, phylogenetic structure reveals geographic and epidemiological patterns of strain distribution, shedding light on global dissemination, local transmission dynamics, and population expansion events. Together, these insights underscore the practical value of phylogenetic approaches for surveillance, risk assessment, and the interpretation of clinical and experimental findings in these important opportunistic pathogens.

This thesis investigates the phylogenetic diversity and population structure of *C. albicans* and *N. glabratus*, two medically important fungal pathogens. Both species are sufficiently distinct phylogenetically yet share similar traits in their ability to survive within and outside a host environment. This work addresses gaps in our understanding of global, regional, and intra-host population diversity through large-scale genomic and phylogenomic analyses, including diversity within hosts with a history of RVVC.

1.6.1 Chapter Two

1. Does WGS-based haplotype analysis support existing clade designations in *C. albicans*, or does it reveal evidence of novel or misclassified clades?
2. How can clades be defined using a more consistent and reproducible approach that also facilitates the recognition of novel clades?
3. What is the current phylogeographic structure of *C. albicans*, based on globally distributed WGS isolates from diverse sources of isolation?
4. What is the phylogenetic diversity of *C. albicans* in Manitoba?
5. What is the distribution of aneuploidy, heterozygosity, and MAT locus types across the species?
6. How distinct are an increased set of environmental isolates from human isolates?

Hypotheses: *C. albicans* harbors greater global genetic diversity than previously described. By integrating whole-genome sequencing with haplotype-aware variant calling and a clade assignment framework, we will improve the reproducibility of population structure and identification of novel or misclassified clades. Furthermore, analysis of aneuploidy, heterozygosity, and MAT locus variation across globally distributed isolates (including both human and environmental sources) will reveal specific genomic patterns. Regional investigations, such as of isolates from Manitoba, will contribute to a more complete understanding of local diversity within the global population structure.

1.6.2 Chapter Three

1. Do whole-genome sequencing-based phylogenetic clusters correspond consistently with multilocus sequence typing sequence types?
2. How prevalent is genomic admixture among *N. glabratus* isolates?

Hypothesis: Most WGS-defined phylogenetic clusters will correspond to a single MLST sequence type. Genomic analysis will identify admixed isolates within some sequence types.

1.6.3 Chapter Four

1. What is the level of standing genetic variation within vaginal and rectal yeast populations in individuals with RVVC?
2. How many isolates per individual are needed to accurately capture population-level genetic diversity in clonally reproducing yeasts like *C. albicans* and *N. glabratus*?

Hypothesis: In individuals with RVVC, vaginal and rectal yeast populations exhibit similar standing genetic variation due to frequent intra-host migration.

Chapter 2: An updated whole genome sequencing phylogeny and genomic diversity of a global collection of *Candida albicans* isolates

This chapter will be submitted to a peer-reviewed journal article.

Contributing authors: Adamu Bukari, A. R., Cuomo, C., Hittinger, C., Opulente D., and Gerstein, A. C.

2.1 Abstract

Candida albicans is a common commensal yeast and opportunistic pathogen characterized by extensive ecological and genetic diversity. Previous phylogenomic analyses have outlined major clades, but these studies often relied on unphased genomes, limited geographic and ecological sampling, and a phylogenetic resolution strategy that has not been universally standardized within the research community. Here, we analyzed phased genomes from 938 isolates and applied a threshold-based clustering approach threshold-based clustering method agnostic to prior clade designations to systematically define clade boundaries while examining genomic features such as aneuploidy, mating-type locus (MTL), heterozygosity, and RNA interference (RNAi) disruption. Our analyses preserved previously defined clades while revealing additional structure, including six novel clades. Clade membership was significantly associated with both geographic origin and isolation source. Compared to isolates from other niches, isolates from the circulatory system exhibited significantly higher heterozygosity, suggesting increased genomic variability in systemic infections. Clade-specific heterozygosity and loss-of-heterozygosity patterns revealed divergent evolutionary trajectories. Most isolates were heterozygous at the MTL, although homozygous forms were enriched in some clades, potentially enabling mating and recombination. Analysis of the *Ago1 PAZ* domain revealed both known and novel RNAi-disruptive variants, predominantly in a heterozygous state. Aneuploidy, present in 8% of isolates, varied by clade and chromosome. Intra-host analysis revealed predominantly clonal colonization, though a subset of individuals harbored multiple clades. This study refines the phylogenetic structure of *C. albicans*, demonstrating how genomic features such as heterozygosity, MTL

composition, and RNAi disruption vary across clades and provide insights into the pathogen's ecological adaptation and potential for genomic plasticity.

2.2 Introduction

Candida albicans is a common human fungal opportunistic pathogen (Brown et al., 2012; Friedman and Schwartz, 2019; Ruhnke, 2006). It is responsible for a wide range of conditions, from superficial mucosal infections, such as oropharyngeal, oesophageal and vulvovaginal candidiasis (Pankhurst, 2013; van Schalkwyk et al., 2015; Vila et al., 2020) to life-threatening systemic candidiasis infections of the blood, heart, central nervous system, eyes and other internal organs (Jampol et al., 1996; Kitaya et al., 2023; O'Brien et al., 2011; Pappas et al., 2016; Puius and Scully, 2007; Turgut et al., 2019). The mortality rates associated with invasive candidiasis have been reported to range from 10% to 47% (Brown et al., 2012; Pappas et al., 2018), and the annual cost of treating candidemia in the US exceeds $1 billion (Miller et al., 2001). *C. albicans* is also commonly found as a commensal species in many of the same body sites where it causes disease (Drell et al., 2013; Ghannoum et al., 2010; Kashem and Kaplan, 2016; Nash et al., 2017). The switch to pathogenicity is known to occur when the local microbiota is disrupted, tissue barriers are compromised, or the immune defenses are weakened (d'Enfert et al., 2021). However, pathogenicity can also occur idiopathically without any obvious predisposing factors in common forms of disease, such as vulvovaginal candidiasis (Sobel, 2016). *C. albican*s has been classified as a critical priority pathogen by the World Health Organization (WHO, 2022) due to global prevalence and increasing rates of drug resistance.

The *C. albicans* diploid genome is about 14.5 Mb and encodes ~ 6,000 genes (Muzzey et al., 2013). *C. albicans* is predominantly asexual and hence its genome is released from the karyotypic constraint imposed by frequent meiosis. In addition to "standard" single-nucleotide variation, small copy number changes, and loss of heterozygosity (LOH),

chromosomal aneuploidy is also frequently observed in lab experiments and among clinical isolates (Forche et al., 2018; Smith et al., 2022; Sui et al., 2020). It is often hypothesized that aneuploidy, which can occur at a higher mutation rate than the per-base pair mutation rate, enhances the ability of *C. albicans* to persist and thrive in the wide range of environments it encounters in the human body (d'Enfert et al., 2021; Mayer et al., 2013). Epigenetic variation, enabling rapid acclimation, likely also plays a role in the wide niche breadth. RNAi was long thought to be inactive in *C. albicans* since the reference strain, SC5314, carries an inactivating homozygous missense mutation in the PAZ domain, which encodes the central RNAi component. However, a recent assessment of 295 additional isolates found that only eight isolates carry the missense variant (Iracane et al., 2024). Of these, seven variants are heterozygous and likely retain RNAi activity, while only one additional isolate was homozygous for SC5314, the inactive variant.

The phylogeny that is widely used for *C. albicans* was published by Ropars *et al.* in 2018 with short-read whole genome sequencing data from 182 global isolates (Ropars et al., 2018); we refer to this as the "existing WGS tree". They designated clusters based on an earlier phylogeny published built from a multi-locus sequence typing (MLST) scheme of seven genes (*AAT1a, ACC1, ADP1, MPIb, SYA1, VPS13,* and *ZWF1b*) using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) to group 1391 isolates (Odds et al., 2007). The MLST phylogeny identified 18 clusters using a P-distance cutoff (the proportion of nucleotide positions exhibiting polymorphism) of 0.04. This cut-off was chosen because it distinguished two clusters containing isolates known to be in separate clades (Clades II and SA) that had previously been identified even earlier using DNA fingerprinting with the moderately repetitive *Ca3* probe (Pujol et al., 2002; Soll and Pujol, 2003). As stated in the

MLST manuscript, the 0.04 cut-off was somewhat arbitrary, used mainly for the convenience of comparative isolate analyses, yet all subsequent phylogenetic studies using updated genome sequencing techniques have adhered to it seemingly without reevaluation (Bensasson et al., 2019; Ropars et al., 2018; Szarvas et al., 2021). The existing WGS tree assigned isolates to 12 numbered clades, matching the previous MLST clade assignments, and five alphabetized clusters (A to E) for novel clusters that contained fewer than ten isolates.

Identifying genetic clusters within pathogenic species - such as the phylogenetic groups underlying population expansions in *Cryptococcus neoformans* (Ashton et al., 2019), the global transmission waves of *Vibrio cholerae* (Mutreja et al., 2011), or the clade-specific antifungal resistance patterns in *Candida auris* (Chow et al., 2020) - is critical for understanding transmission and evolution of these organisms. In some fungal microbial species, isolates within the same clade tend to have similar geographic origins, virulence, and/or resistance patterns, as seen in *Candida auris* (Chow et al., 2020), *Saccharomyces cerevisiae* (Peter et al., 2018) and *Aspergillus fumigatus* (He et al., 2024; Rhodes et al., 2022). Although no definitive geographical link has previously been found with the *C. albicans* clade, in some cases, a large proportion of isolates within a cluster come from broadly identifiable geographic regions. By example, recent MLST analyses of vulvovaginal isolates from northern China revealed a novel cluster of 92 *C. albicans* isolates (Song et al., 2022). Thirteen (13/46) isolates from Thailand were identified to cluster together in clade 17 (Pham et al., 2019). It stands to reason that any potential geographical specificity of clades is gradually being lost, perhaps due to the high global rate of human movement, which facilitates not just transmission of isolates but also increases the possibilities of recombination events

producing novel lineages (Odds et al., 2007; Odds and Jacobsen, 2008; Song et al., 2022). *C. albicans* may be particularly susceptible to this due to its high colonization prevalence in healthy hosts compared to other fungal microbes (such as *Trichophyton* spp and *Pneumocystis* spp*.)* which are mostly obligate pathogens.

Many studies have sought to determine whether there is a link between clade and phenotypic traits of clinical interest. Clade 1 isolates are the most likely to be resistant to flucytosine, achieved through the R101C mutation in *FUR1* that is not observed in flucytosine-resistant isolates from other clades (Dodgson et al., 2004; Odds, 2009; Pujol et al., 2004). Clade 1 isolates are also the most likely to be resistant to terbinafine compared to clade 2, 3, 4 and 11 isolates (Odds, 2009), yet no difference was observed for a panel of seven other common antifungal drugs. Isolates in clade 2 have significantly lower levels of acid phosphatase activity than clade 1 and 3 isolates (MacCallum et al., 2009). Significant differences among isolates in different clades (1, 2, 3, 4) in relation to midrepeat sequence alleles of gene families that encode *C. albicans* surface proteins that play a role in adhesion to host surfaces (*ALS2*, *ALS4*, *ALS6*, *ALS7*, *ALS9*, *HYR1*, and *HYR2*) have also been identified (MacCallum et al., 2009). Blood isolates from clade 17 had greater hemolytic activity compared to isolates from eight other clusters, though no differences were found for proteinase activity, phospholipase activity, or biofilm formation. However, no clade association was observed in terms of biofilm formation, growth in bovine serum albumin, growth in different temperatures and adherence to plastic catheter (MacCallum et al., 2009). A significantly higher proportion of isolates from clade 1 are associated with superficial infections and commensal carriage compared to isolates from other clades, and clade 1 is repeatedly the largest clade, regardless of the phylogenetic method used (Odds et al., 2007;

Odds and Jacobsen, 2008). It has been hypothesized that clade 1 isolates may have an enhanced ability to evade host defenses due to the possession of a 985 bp *HpaII* restriction enzyme fragment (*MU13-4*), which potentially has a role of assisting strains in generating genetic variability. Whether the identified associations have clinical relevance in a predictive manner that could inform treatment decisions remains an open area of study (Giblin et al., 2001).

The existing WGS tree is composed of 160 isolates acquired from a wide variety of human sites, 19 isolates from food spoilage, two isolates from the feces of *Sturnus vulgaris* and an isolate of unknown source (Ropars et al., 2018). The set is thus very biased towards isolates from or associated with humans. Yet *C. albicans* has been isolated from animals (Talazadeh et al., 2022) and diverse ecological environments, including beach water and sand, oak bark, and soil (Bensasson et al., 2019; Hamlin et al., 2019; Maciel et al., 2019; Opulente et al., 2019). Very few studies have compared environmental isolates to clinical and human isolates. Using the existing tree as the backbone, Bensasson *et al*. sequenced three *C. albicans* isolates from oak tree bark, finding that two were closely related to human isolates, while the third was not assigned to an existing cluster (Bensasson et al., 2019). The three environmental isolates had a slightly higher genome-wide heterozygosity than the clinical isolates (Bensasson et al., 2019). The transmission dynamics of *C. albicans* remain unclear - although it has often been stated that humans are colonized at birth (Caramalac et al., 2007; Filippidi et al., 2014), there is little to no data tracking strain diversity over time (Ward et al., 2018), and the initial strain present at birth may not persist throughout a lifetime. A much larger panel of environmental isolates is needed to properly assess whether diversity among

environmental isolates differs compared to isolates from other sources and to assess the relationship between environmental and clinical isolates.

Despite the availability of a phased diploid genome for *C. albic*ans since in 2013 (Muzzey et al., 2013), all WGS phylogenetic studies conducted so far have only used variants from the "A" haplotype (hapA of A21 or A22 reference genomes). The current variant calling tools in high use do not specifically account for phased genomes, i.e., to distinguish between the parental chromosome copies. This results in the loss of haplotype information, which could be valuable for accurately constructing phylogenies. The *C. albicans* genome has 4.59 to 8.62 heterozygous variants/kb (Mixão and Gabaldón, 2020), substantially higher than species such as *C. tropicalis* (2-6 heterozygous variants/kb, O'Brien et al., 2021). In phylogenetic analyses, as in the existing WGS tree, unphased heterozygous sites are often encoded using IUPAC ambiguity codes in consensus sequences. This can lead to phylogenies that disproportionately reflect homozygous variation, potentially underestimating genetic diversity.

To improve phylogenetic resolution, address the aforementioned limitations of clade designation, and assess potential geographic signatures in we reconstructed a new short-read WGS phylogeny for *C. albicans*. We used data from 1178 isolates, including 1130 human-associated isolates, 31 environmental isolates, and 17 isolates of unknown origin. Among these, 85 are newly sequenced human-associated isolates from Manitoba, Canada, and seven are environmental isolates from the United States, with the remainder sourced from the NCBI SRA archive. We incorporated phased haplotype information and used a statistical threshold-based method for clade assignment. This expanded and rigorously analyzed dataset updates our understanding of the *C. albicans* phylogenetic structure, including the

geographic and site-specific distribution of genome-wide heterozygosity, aneuploidies, and RNAi-deficiencies. We also highlight how biased sampling has led to continuing knowledge gaps in the relationships among isolates between and within populations.

2.3 Materials and methods

2.3.1 Acquisition of sequence data

The NCBI SRA database (Sayers et al., 2022) was searched in July 2024 for all deposited Illumina-sequenced *C. albicans* whole genome sequences. Isolates from experimental evolution studies were excluded. Additionally, only samples with ≥ 80% of reads mapping to the reference genome, breadth of coverage ≥ 80% and mean depth of coverage ≥ 30× were considered (following human WGS analyses; (Ajay et al., 2011; Mardis, 2012)). In total, 1088 isolates from 22 publications were retained (Adamu Bukari et al., 2025; Alkhars et al., 2024; Anderson et al., 2023; Bensasson et al., 2019; Cavalieri et al., 2017; Chew et al., 2021, 2023; Cuomo et al., 2019; Ford et al., 2015; Gnaien et al., 2024; Gong et al., 2023; Guinea et al., 2021; Hirakawa et al., 2015; Li et al., 2022; McTaggart et al., 2020; Mohammadi et al., 2023; Peterson et al., 2023; Ropars et al., 2018; Sitterlé et al., 2019, 2020; Szarvas et al., 2021; Zuber et al., 2023). In addition, we included fastq data from seven newly sequenced environmental isolates, and 83 newly sequenced isolates acquired in 2012 and 2018 from the Health Sciences Centre Microbiology Lab in Manitoba (with assistance from the Shared Health/Diagnostic Services), Canada (Table S2.1 at https://github.com/microstatslab/Calbians_phylogenetics/supporting_tables) for a total of 1178 isolates ("complete isolate set"). The isolate collection included 361 isolates from 95 individuals with more than one isolate: 71 people with isolates taken from multiple timepoints, 24 people with multiple isolates taken simultaneously from the same timepoint (either at the same or different body sites, Table S2.1).

2.3.2 DNA extraction and sequencing of Manitoba isolates

Genomic DNA was extracted from single colonies of 83 Manitoba isolates using a standard phenol-chloroform protocol previously described (Kukurudz et al., 2022). DNA quality and concentration were assessed spectrophotometrically (NanoDrop 2000, Thermo Scientific™) and fluorometrically (Qubit® 2.0 Fluorometer with the dsDNA BR Assay Kit, Invitrogen™), respectively. The genomes of the 43 isolates from 2012 were sequenced by the Microbial Genome Sequencing Center (Pittsburgh, USA) using the Illumina NextSeq 550 sequencing technology with paired-end reads of 150 bp. The bcl-convert v3.9.3 package (https://support-docs.illumina.com/SW/BCL_Convert/Content/SW/FrontPages/BCL_Convert.htm) was used in demultiplexing, quality control, and adapter trimming. The genomes of 40 isolates from 2018 were sequenced by the Genome Quebec Innovation Center in Montreal using NovaSeq6000 S4 sequencing technology with paired-end reads of 150 bp. Reads have been deposited at the National Center for Biotechnology Information (NCBI) Sequence Read Archive under BioProject ID PRJNA991137.

2.3.3 Read mapping and variant calling

The reads from all isolates were visually inspected for quality with FastQC v0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed and filtered with Trimmomatic v0.36 (Bolger et al., 2014). Read mapping and variant calling were performed using HaploTypo v1.0.1 (Pegueroles et al., 2020) using the default parameters. Briefly, HaploTypo is run in four successive modules (see scripts in https://github.com/Gabaldonlab/haplotypo): read mapping (mapping.py), variant calling (var_calling.py), inference of true alternative variants for each haplotype

(VCFcorr_alleles.py) and reconstruction of phased haplotypes (haplomaker.py). Filtered reads were mapped with readgroups added onto the phased haplotypes (hapA and hapB) of the A21 SC5314 reference genome (Muzzey et al., 2013) with BWA-MEM v0.7.18 (Li, 2013). Picard v3.1.0 (http://broadinstitute.github.io/picard) was used to sort the alignment, convert the SAM alignment to BAM format, and mark duplicate reads. Alignment quality was assessed with the tool CollectAlignmentSummaryMetrics from picard v3.1.0 (http://broadinstitute.github.io/picard) and consolidated across all samples with MultiQC (Ewels et al., 2016). Average depth of coverage was assessed with samtools coverage (Danecek et al., 2021) of samtools v1.20. Bcftools v1.19 (Danecek et al., 2021) was used for calling and filtering variants using the default parameters in Haplotypo. The VCFcorr_alleles.py script of Haplotypo was used to compare the variant calling results from each sample against each of the two phased reference haplotypes to generate two VCF files, one for each haplotype, reporting the variants specifically observed in each of them while ignoring ambiguous genotypes (-amb 0).

2.3.4 Comparison of trees constructed with and without haplotype information

To determine if considering haplotype information affected tree topology, we constructed a phylogeny using the 148 isolates from Ropars *et al.* (2018) (i.e., the existing WGS phylogeny) as there exists clade information for these isolates. We retained only one of the 35 clade 13 isolates (i.e 148/182) for rooting the phylogenies. We generated a phylogeny from the hapA (this is the main haplotype used in all *C. albicans* phylogeny analyses) intermediate vcf files from running HaploTypo. We also generated a phylogeny from the MLST alleles (*AAT1a, ACC1, ADP1, MPIb, SYA1, VPS13,* and *ZWF1b*) with loci extracted from

the same intermediate vcf files. We then generated a phylogeny from the final consensus fasta files generated from considering haplotype information as detailed above.

To compare the trees, the pairwise distance between tree topologies was assessed using RF.dist and KF.dist functions from the phangorn package (Schliep, 2011) in the R Programming lanugage. The normalized Robinson-Foulds (nRF) distance and branch score distance (KF) were calculated; nRF distance reflects the number of bipartitions differing between topologies, whereas the KF distance quantifies the difference in branch lengths and tree topology between the trees (01_Tree-topology_comparison.R). Therefore, two identical topologies will receive a value equal to 0 with both metrics. Conversely, distance values will increase (to max to 1 in nRF) as the compared trees become more different.

2.3.5 Phylogeny construction from the complete isolate set

A FASTA file for the reconstructed haplotype from each isolate was generated using the haplomaker.py of haplotypo script. This process ensures that heterozygous positions are not disregarded or replaced by IUPAC ambiguity codes, as is observed in many pipelines (Ortiz, 2019). The two fasta sequences for each chromosome from each isolate sample were concatenated into a single sequence, and then all isolate sequences were combined into a single multiple sequence alignment file. This file was input to FastTree (v2.1.11, Price et al., 2010) in the double precision mode to construct a maximum-likelihood phylogenetic tree using the general time reversible model and the -gamma option to rescale the branch lengths. FastTree has been found to produce equally accurate trees with large datasets as other ML-based phylogeny predictors such as RAxML (Stamatakis, 2014) within a significantly shorter time (Liu et al., 2011). The resulting phylogeny was visualized and annotated with the

Interactive Tree Of Life (iTOL, v5; Letunic and Bork, 2021). We used the 37 *C. africana* ("clade 13") isolates to root the tree (Mixão et al., 2021; Romeo et al., 2013)

We sought to avoid potential bias in downstream analyses from the 361 isolates coming from the 95 individuals who were sampled multiple times ("intrapopulation isolates"). The majority of intrapopulation isolates are from the same body site and clustered monophyletically together. In these cases, we randomly picked a single isolate to retain from each monophyletic group from each individual. In addition, we retained intrapopulation isolates that were not monophyletic and isolates from different body sites. We thus retained 128 of the 361 intrapopulation isolates. We generated a new phylogeny of 945 of these "phylogenetically informative" isolates, including the 37 *C. africana* isolates used to root the phylogeny.

2.3.6 Clade delineation

To delineate the clades within the phylogeny, we ran TreeCluster (Balaban et al., 2019). TreeCluster uses several functions to agnostically identify clusters within phylogenetic trees. We selected methods that are optimized to identify clades within the phylogeny (i.e., methods that have the “clade” suffix). This included max clade, which is the default method. We additionally examined the single linkage method as a representation of the three single linkage methods (single linkage cut, single linkage union). To identify a statistically well-supported phylogeny, we ran each method with threshold values from t = 0-1, increasing t by 0.001, for a total of 1000 values per method. We identified regions of parameter space where a range of threshold values yielded the same number of clades. We

then visually inspected each clade assignment on the phylogeny using iTOL and compared it to the existing WGS phylogeny.

2.3.7 Heterozygosity analyses

Sites with missing data are common with short-read WGS data and can significantly impact heterozygosity estimates. We thus excluded isolates with > 15% missing genotype positions, retaining 745 isolates from the phylogenetically informative isolate set. The average heterozygosities of the isolates were calculated and statistically compared by isolation site and clade after prior testing of the homogeneity of variances assumption with Levene's test.

Genome-wide heterozygosity at each was calculated using PLINK (v2.0), a widely used tool for genetic data analysis (Purcell et al., 2007). Bcftools was used to call variants in the consensus mode to ensure all sites are considered using the A21 hapA reference (Muzzey et al., 2013). Genotype counts for each isolate were obtained by extracting the number of homozygous reference (hom), homozygous alternate (homalt), and heterozygous (het) genotypes from the VCF files. To calculate the genome wide heterozygosity across the isolates, *bcftools query* (options -H and -f) was used to generate a table of genotypes in all positions of 738 isolates. A custom script (06-heterozygosity_analyses.R) was used to calculate and visualize the heterozygosity for 5kb sliding window across the genome. Centromeric and subtelomeric regions (defined as 15 kb from the start and end of each chromosome) were excluded as they are known to be prone to artifactual errors in short read data due to repeats.

2.3.8 Aneuploidy Analyses

We quantified aneuploidy and copy number variation (CNV) (duplications or deletions of smaller genomic segments) in the complete genome isolate set. Sequencing reads were aligned to the reference genome using BWA-MEM v0.7.18 as detailed above. Post-alignment, we used samtools depth to calculate the depth of coverage at each genomic position from each BAM file., generating a comprehensive coverage profile across all chromosomes. Coverage data was processed and visualized using a custom R script (05a-Aneuploidy.R). The script calculates the average read depth within non-overlapping 5 kb bins across the genome. The median number of reads per chromosome is calculated and used to normalize read depth across bins. Chromosomal aneuploidies and small regions of elevated copy number (CNVs) were visually identified by two people independently, based on the generated plots. Where coverage across at least one chromosome (or chromosome part) was a non-integer number, read depth was recalculated to have a base ploidy of triploid or tetraploid and again scored visually. CNV breakpoints were manually determined by determining the bin where read depth changed, which is typically very clear. Example graphs used for quantification are provided in Figure S2.1.

2.3.9 Mating-type loci analyses

The mating-type loci (MTL) in the 938 isolates was identified by aligning sequencing reads to both haplotypes of chromosome 5 from the A22 reference genome (A22-s08-m01-r09) using bcftools. Samtools depth was employed to calculate both the depth and breadth of coverage across the MTLa region (Ca22chr5A_C_albicans_SC5314:393493–394455 and

394560–395220) and the MTLα region (Ca22chr5B_C_albicans_SC5314:395642–396223 and 401608–402227). Loci with coverage < 0.5 were considered inactive or absent.

2.3.10 Identifying variants in the AGO1 gene

We sought to quantify the prevalence of mutations in the AGO1 PAZ domain, linked to RNAi activity, in the 907 isolate set. A BLAST search of the gene was used to pinpoint the exact coordinates of the PAZ domain in the A21 *C. albicans* chromosome 4 reference (i.e chr4_A:1408039-1408347). The corresponding region was then extracted from the vcf files from all isolates using *bcftools filter* and converted to a multiple sequence alignment file. Missing data were assumed to be homozygous consensus. For heterozygous regions, the homozygous and alternate alleles were independently incorporated into the reference to construct the haplotype sequences for each isolate to identify unique PAZ domain sequences. The nucleotide sequences were aligned and converted to amino acid sequences using Clustal Omega in SnapGene, with codon table 12. The frequencies of unique PAZ domain sequences were then calculated, and the regions of differences were identified.

2.3.11 Analysis and visualization reproducibility

All data, supplementary tables and code required to reproduce all statistical analyses and visualizations (with the exception of the phylogenies) are available at

https://github.com/microstatslab/Calbians_phylogenetics

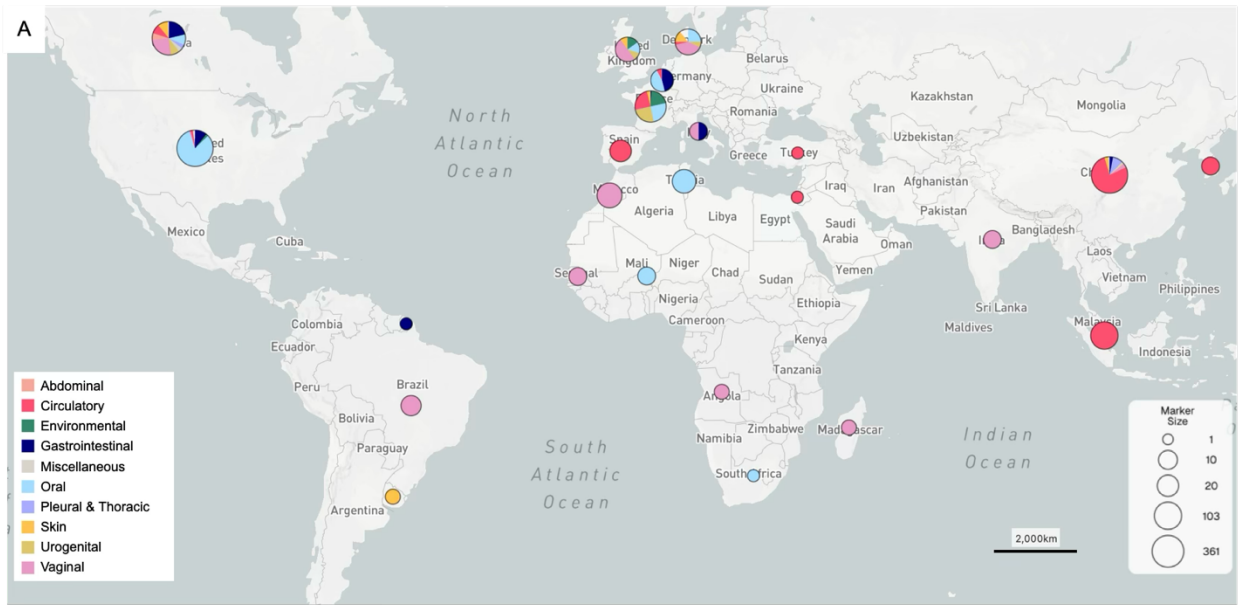
Supporting tables can be found at

https://github.com/microstatslab/Calbians_phylogenetics/supporting_tables

2.4 Results

2.4.1 Isolate information

The 1178 isolates with WGS data included in this study were obtained from 26 countries across five continents. The distribution across countries and continents was very uneven (Figure 1); the continents with the highest representation are North America (520 isolates) and Asia (410 isolates), followed by Europe (179), Africa (51) and South America (10). We identified no isolates from Australia. At the country level, it is even more patchy, with only eleven countries having at least than 10 isolates (Table S2.1: Belgium, Canada, China, Denmark, France, Morocco, Singapore, Spain, Tunisia, United Kingdom, USA). We grouped the isolate into ten isolation source categories (Figure 2.1, Table S2.1). The majority of isolates from each country come from a single isolation source; only eight countries have isolates from multiple sources. Isolates from France and Canada are the most diverse by source; the majority of Canadian isolates were sequenced for this project and were collected at the Health Sciences Centre in Winnipeg. Approximately one-third of all isolates were from the circulatory system (mainly blood) , and another third from the oral cavity. The least common sources are abdominal and environmental. The environmental isolates include 19 from food spoilage (from France), two from birds (from France), and ten from plants or soil (three from oak trees in the UK, and seven newly sequenced for this project from soil in the US). This uneven distribution is likely influenced by many factors such as regional research focus and funding availability. Thus, although this study represents the largest global survey to date for *C. albicans*, there are very likely to be biologically relevant regional differences among isolates across the globe which are not all captured here.



B

(69) Adamu Bukari *et al.* 2023 *bioRxiv*
(217) Alkhars *et al.* 2024 *PLoS One*
(31) Anderson *et al.* 2023 *PLoS Biol*
(3) Bensasson *et al.* 2019 *Genetics*
(2) Cavalieri *et al.* 2017 *Front Immunol*
(4) Chew *et al.* 2021 *Lett Appl Microbiol*
(40) Chew *et al.* 2023 *Microb Genom*
(1) *Cuomo et al. 2019 Microb Res Ann*
(37) Ford et al. 2015 *eLife*
(16) Gnaien *et al.* 2024. *PLoS Pathog*
(343) Gong *et al.* 2023 *Microbiol Spectr*
(10) Guinea *et al.* 2021 *Med Mycol*
(18) Hirakawa *et al.* 2015 *Genome Res*
(26) Li *et al.* 2022 *Nature*
(5) McTaggart *et al.* 2020 *AAC*
(12) Mohammadi *et al.* 2023 *BMC Genomics*
(1) Peterson *et al.* 2023 *Genomics*
(175) Ropars *et al.* 2018 *Nat Commun*
(16) Sitterlé *et al.* 2019 *Sci Rep*
(8) Sitterlé *et al.* 2020 *J Antimicrob Chemother*
(28) Szarvas *et al.* 2021 *J Fungi*
(1) Zuber *et al.* 2023 *Microorganisms*
(25) Not published
(90) This study

Figure 2.1: Distribution and literature source of isolates used in this study. A) Geographical distribution of the 1178 isolates analyzed in this study. Circle size is proportional to the log of the number of isolates from each region, while colored sectors represent their respective isolation sources. This includes 37 clade 13 (*C. africana*) isolates. The image was generated using Microreact. B) Publication source of the genome sequences used; the number of isolates from each manuscript is provided in brackets.

2.4.2 Topological differences in constructing phylogenies with haplotype information

A primary goal of this project was to create an updated intraspecific phylogeny for *C. albicans*. The existing WGS tree included short-read, unphased data from 182 isolates. To assess the impact of incorporating haplotype phasing information, we reconstructed this tree ("unphased WGS") and compared it to trees with the same isolates with haplotype phasing ("phased WGS") and a multilocus sequence typing tree (MLST). The normalized Robinson-Foulds (nRF) distance between the phased and unphased WGS trees was 0.655.

The nRF distance between the phased WGS tree and the MLST tree was even higher (0.862). Combined, this demonstrates that including knowledge of haplotype information does influence the tree topology, which can also be seen by visually comparing the trees (Figure 2.2). Although the MLST tree does partially recapitulate the WGS tree and overall genetic divergence patterns, it lacks the phylogenetic resolution provided by genome-wide data, suggesting MLST analysis should be used cautiously if precise phylogenetic relationships are desired.

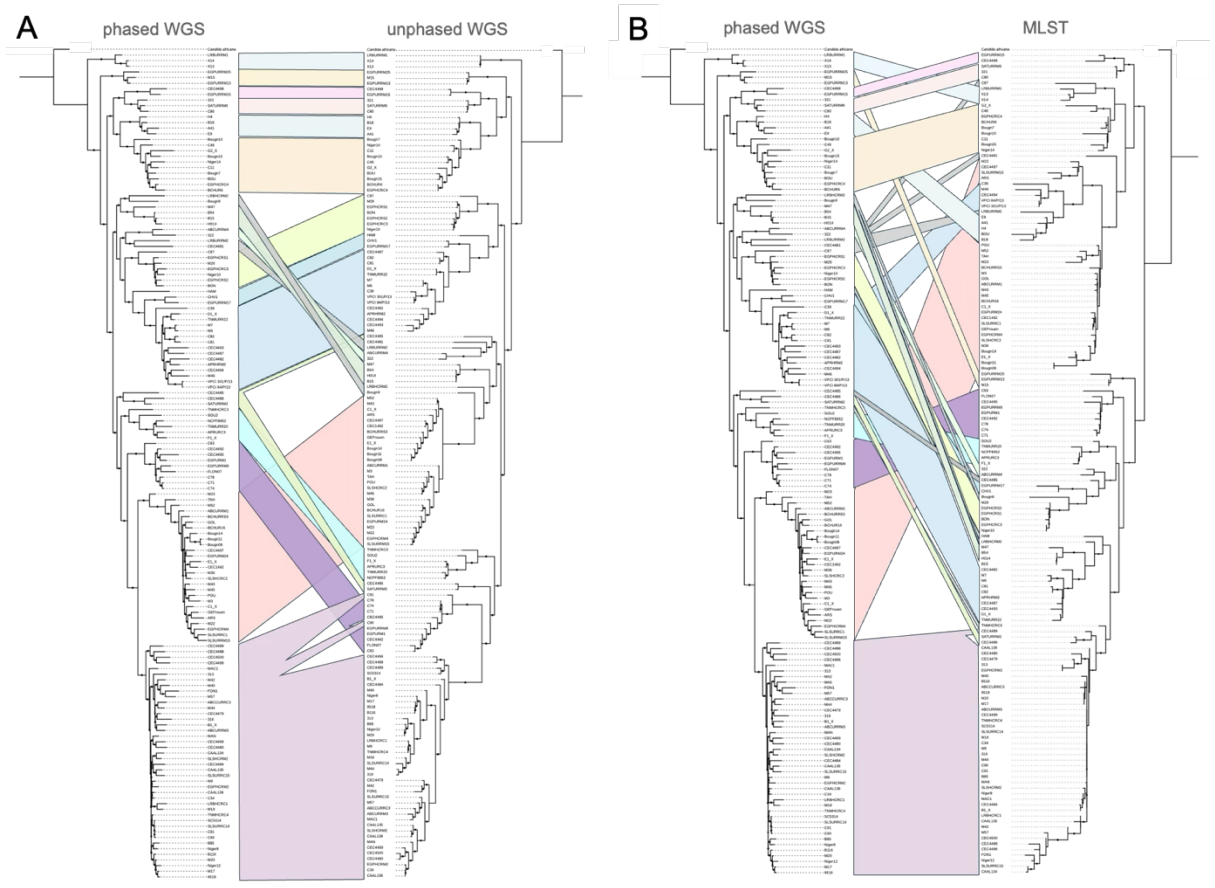


Figure 2.2: Topological differences in within-species phylogenies created with different sequence analysis methods. On the left side of both panels is phased short read WGS data. This is compared to A) a phylogeny constructed from unphased short read WGS data, and B) a phylogeny constructed from MLST sequences. The color bands in both cases indicate a match in the order of isolates and were drawn to assist with a visual interpretation of the differences.

2.4.3 Updated *C. albicans* phylogeny

As haplotype phasing affects both tree topology and branch lengths we therefore constructed the updated global whole-genome phylogeny using phased genomes from the full (1178) isolate set. Once the initial phylogeny was generated, we visually assessed the 361 intra-population isolate set that included multiple isolates from the same person. For each of the 95 people with multiple isolates, we determined the minimum number to retain to reduce monophyletic isolate clusters from the same person; we thus retained 128 intra-individual isolates (Figure S2.2). The majority of removed isolates came from North America (n=217). The 35 isolates that were identified as clade 13 (*C. africana*) from the existing WGS and two additional isolates that grouped with them were included as the outgroup. The final tree is thus comprised of 908 *C. albicans* "phylogenetically informative" isolates (and 37 *C. africana*), which we refer to as the "phylogenetic informative set" of isolates (Table S2.1).

The overall topology of the final phylogeny largely but not entirely recapitulated the existing WGS tree clade structure (Figure S2.3). Five isolates from the previous WGS tree clustered differently: the two clade B isolates clustered with clade 2 isolates, the two clade D isolates clustered with clade 20, one clade 3 isolate clustered with clade 1. In the much larger strain set, all ten singletons had closely related isolates.

To delineate clades for the updated tree, we sought to adopt a statistically-supported approach. The existing tree relied on the somewhat arbitrary p-distance cut off of 0.004 initially implemented on an MLST tree to delineate clades. We implemented seven TreeCluster strategies, which seek to determine phylogenetic tree topologies by clustering tip sequences using different optimization functions and distance thresholds. *A priori*, we were expecting to find a function that would yield approximate 17 clades (the number from

the existing tree) with substantial overlap in clade assignment, given that isolate designations have been relatively consistent over time through different sequencing technologies. Two strategies yielded broad regions of parameter space where the same number of clades was identified (Figure 2.3A). The “max clade” strategy, the default setting, identified 20 clades with threshold parameter space values from 0.012-0.014 while the "single linkage" strategy (which has previously been used in HIV research) identified 24 clades over values 0.0047-0.0055, excluding 0.0051. For comparative purposes, we also visually examined single linkage at value 0.009 which identified 109 clades.

When we visually mapped the clade designations on to the final phylogeny, it was clear that the clades identified from the max clade strategy well matched our *a priori* expectations and the previous WGS tree (Figure 2.3B). Clades from the previous WGS phylogeny were retained (with the expected merge of clade B with clade 2). In addition, six novel clades were identified. To maintain consistency with past nomenclature, we retained the original Arabic numeric clade labels. Alphabetical clade names from the previous WGS phylogeny and novel clades were assigned numeric designations based on their location in an anti-clockwise manner (starting from clade 1), which is somewhat consistent with the existing clade names and locations. Given that four MLST clades (5, 6, 7, 14) are absent in the WGS phylogenies we reassigned these numbers. Following this scheme, clade A was assigned to be clade 20, clade C to 15, clade D to 21, clade E to 5, and the five new clades to 6, 7, 14, 17, and 19 (Figure 2.4).

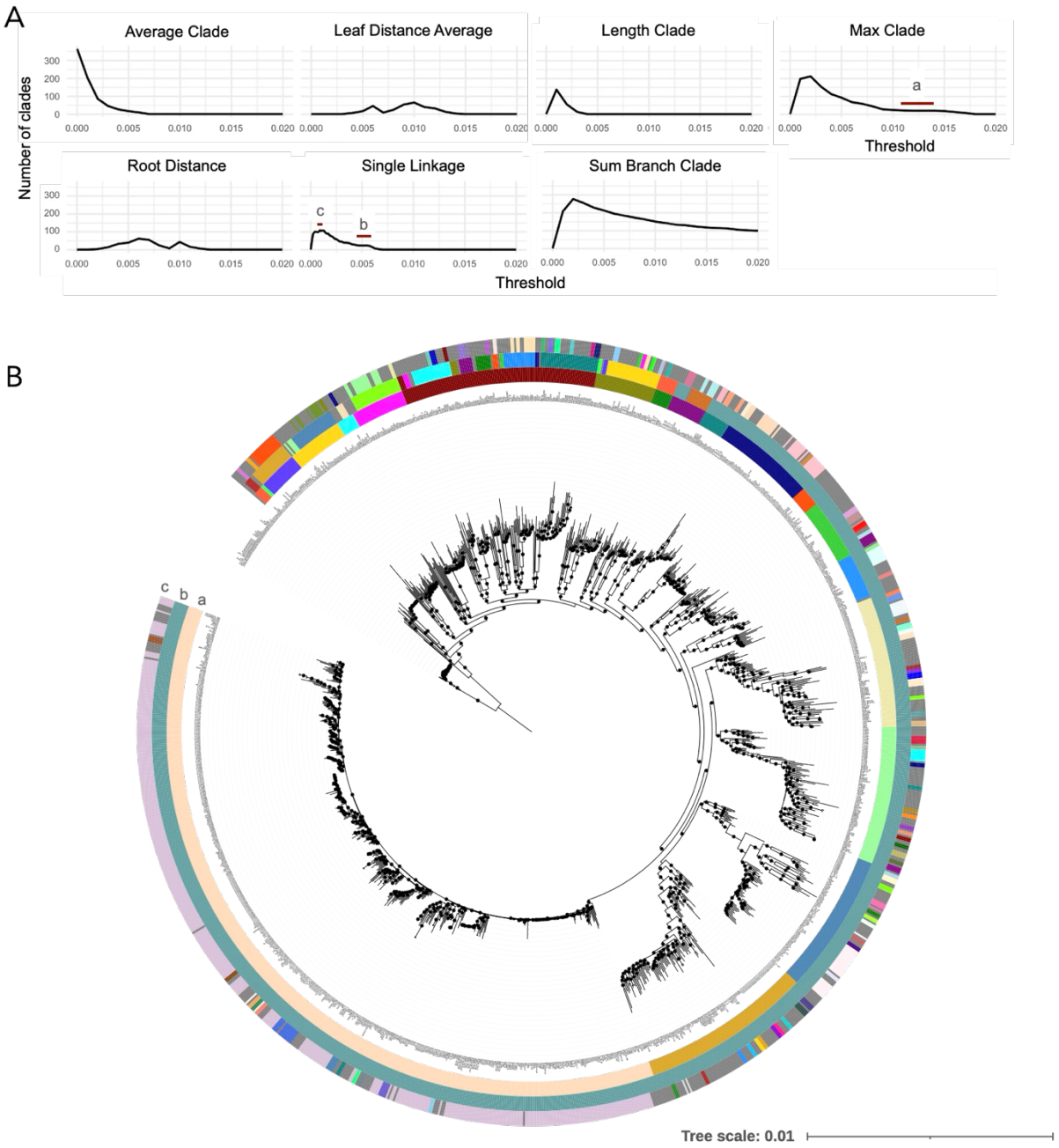


Figure 2.3: Clade delineation using TreeCluster across multiple methods. A) The number of predicted clades across a range of threshold values (0 to 1, in 0.001 increments) using seven TreeCluster methods. Regions with stable clade predictions-defined as consistent clade sizes across multiple consecutive thresholds, are highlighted with a dark brown bar and labelled a–c. These include a range of threshold values in the “max clade” strategy (clade size = 20, thresholds t = 0.012 – 0.014) and two threshold ranges in the "single linkage" strategy (clade sizes = 24 and 109). B) Visualization of the stable clades identified in (A) mapped onto the phylogenetic tree.

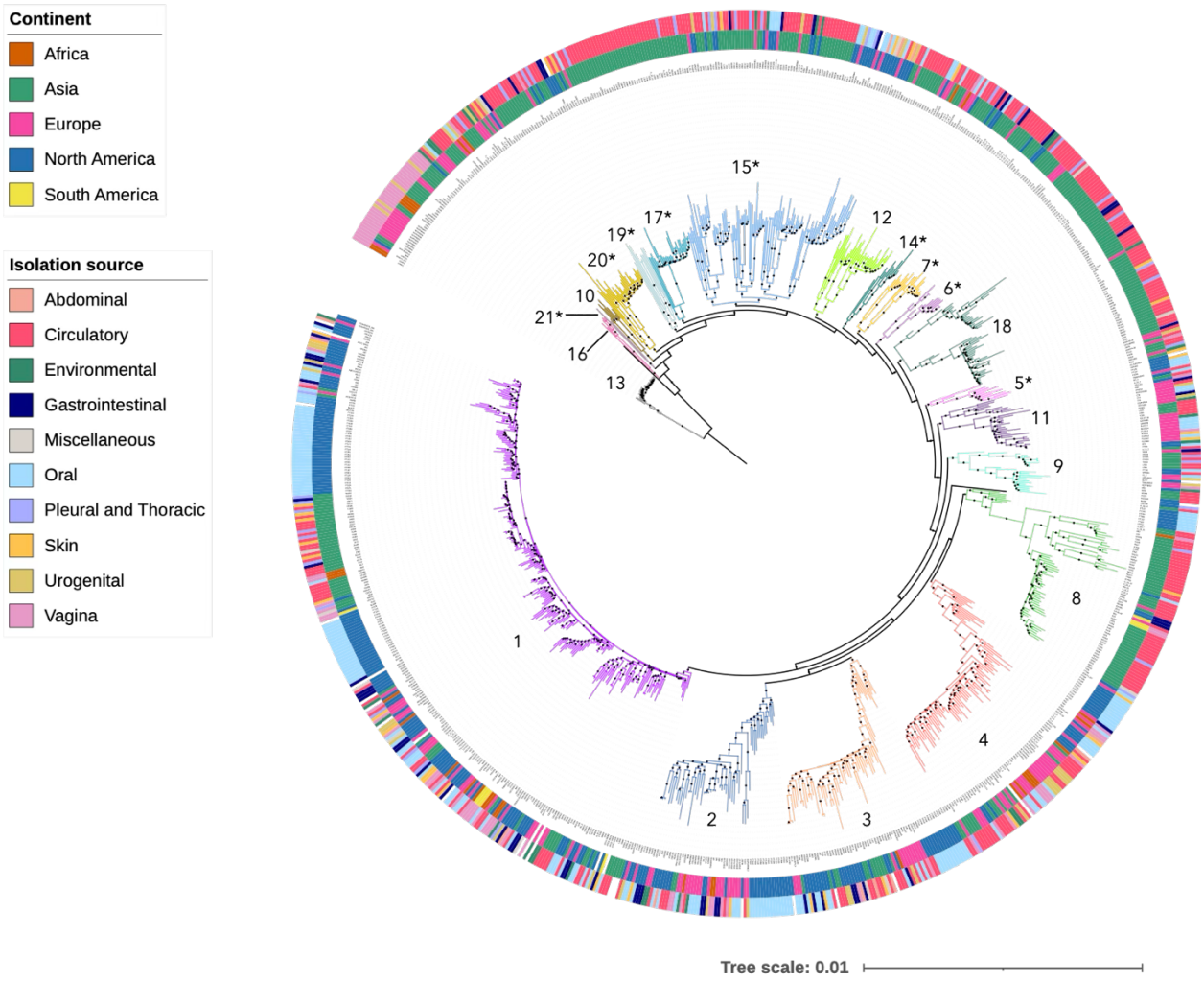


Figure 2.4: Maximum likelihood phylogeny of 938 isolates. Branch colors correspond to clade designations as defined in the study. The clades numbers are denoted on the clades and asterisks are used to show new clade labels. Bootstrap support values ≥80% are indicated by ribbons on the branches. Three concentric rings surround the phylogeny: the innermost ring are strain names, the middle ring represents the continent of origin for each isolate, while the outermost ring denotes the source of isolation. This visualization highlights both the phylogenetic structure and the geographic and ecological diversity across clades.

2.4.4 Geography and site of isolation are nested together

The distribution of isolates in the full phylogeny is relatively similar to the previous phylogeny. The distribution of isolates is not equal among clades, clade 1 in particular contains approximately a quarter of all isolates, while three clades (10, 16, 21) have less than ten isolates.

Clade, isolation source (Figure S2.4 and S2.5), and geographic origin (Figure S2.6) are not independent of each other. This is seen visually on the map (Figure 2.1) and as overlapping color blocks in the outer rings in the phylogeny (e.g., North American oral isolates exhibit several clean blocks in multiple clades; Figure 2.4). The strongest association was observed between isolation source and continent (Cramér's V = 0.55). The largest contributors to this association are isolates from the most common sources. Circulatory isolates are over-represented in Asia relative to North America (and to a degree, Europe), while oral isolates are similarly over-represented in North America compared to Asia (Figure S2.7). Interestingly, Africa and South America are both over-represented for vaginal isolates, while Europe is over-represented for environmental and urogenital isolates (Figure S2.7). Clade and source (V = 0.23) and clade and continent (V = 0.37) are also both moderately associated. Statistically, all three factors have significant pairwise associations (pairwise Chi-square tests with 100000 Monte Carlo simulations: source × continent: $\chi^2 = 1054$, $p < 0.0001$; source × clade: $\chi^2 = 433$, $p < 0.0001$; continent × clade: $\chi^2 = 473$, $p < 0.0001$). These results suggest that while clade distribution shows some significant structure based on geography and source, these factors are largely confounded.

2.4.5 Manitoba as a case study to remove geography

To examine the phylogenetic relationship among isolates from different sources with the potential confounding factor of geography removed, we examined the set of 83 isolates collected from a hospital microbiology lab from Manitoba. This set includes all *C. albicans* isolates collected in 2012 and 2018 from the lab. The isolates fell into nine different clades, eight of which were represented in both years; there was no significant difference in clade composition between the two years (χ^2 = 6.64, p = 0.6534, Figure S2.9). Interestingly, no significant association was observed between clade and isolation source (χ^2 = 66.90, p = 0.3814), and isolates from different sources (and different years) often grouped right beside each other. This highlights, as previously shown, that there is no obvious pattern between clade and isolation source and migration across the globe seems to be relatively common for *C. albicans* genotypes. However, this is incomplete, as a signature of geographic enrichment is present in many clades. Combining the global and local (Manitoba) analyses suggests that geography, rather than isolation source, has more of an impact on shaping phylogenetic relationships.

2.4.6 Karyotypic variation

The full isolate set was examined for aneuploidies and copy number variations (CNVs). 86 isolates exhibited karyotypic variation: 57 isolates had at least one aneuploidy, 30 isolates had at least one CNV region larger than 50 kb, and six isolates had both (Figure 2.5A). Four of these isolates were identified as triploid (3N), and four others as tetraploid (4N), while a base ploidy could not be determined for four isolates. All of the triploid isolates had multiple aneuploid chromosomes, while the tetraploids had only a single. The use of

WGS coverage to determine ploidy precludes identification of euploid polyploids, so this represents a lower limit for ploidy variants in the strain set. There was a significant negative correlation between chromosome size and the number of aneuploidies (Pearson's correlation: t_6 = -2.80, p = 0.031, cor = -0.75), potentially indicating less constraint against aneuploidy on smaller chromosomes that carry fewer genes. The exceptions were chromosomes 3 and 6, which both had fewer aneuploidies that their size would have predicted, suggesting there might be stronger selection against extra copies of these chromosomes. There was no correlation between chromosome size and the number of CNVs (t_6 = 0.11, p = 0.92, nor between the number of aneuploidies and the number of CNVs (t_6 = -0.11 p = 0.91). The majority of CNVs across all chromosomes were terminal, in many cases in both directions (i.e., CNVs that extended to the telomeres), though for most chromosomes a small number of interstitial CNVs were also observed (Figure 2.5B). Some regions seemed more likely to be involved in a CNV (e.g., regions of overlap among many CNVs on chromosome 6 and chromosome R) (Figure 2.5B).

Karyotypic variant isolates were observed throughout the phylogeny and were typically not clustered together (Figure S2.9). At least one isolate from each isolation source was observed to be a karyotype variant (Figure 2.5C), with the number proportional to the total number of isolates from each source. The two most common sources exhibited different patterns. This hints at a potential beneficial association between selection in these different niches (circulatory versus oral) selecting for different chromosomal variants. However, caution in interpretation is needed given the sampling biases. The circulatory isolates come from multiple locations and authored publications, while the oral isolates all come from a

single study done on North American isolates that was focused on drug resistance (Ford et al., 2015).

A

Chr	Size	Aneuploidy	CNV	Total
Chr1	3188548	3	6	9
Chr2	2232035	6	1	7
Chr3	1799407	2	3	5
Chr4	1603443	12	4	16
Chr5	1190928	15	7	22
Chr6	1033530	10	9	19
Chr7	949616	23	6	29
ChrR	2286389	5	18	23

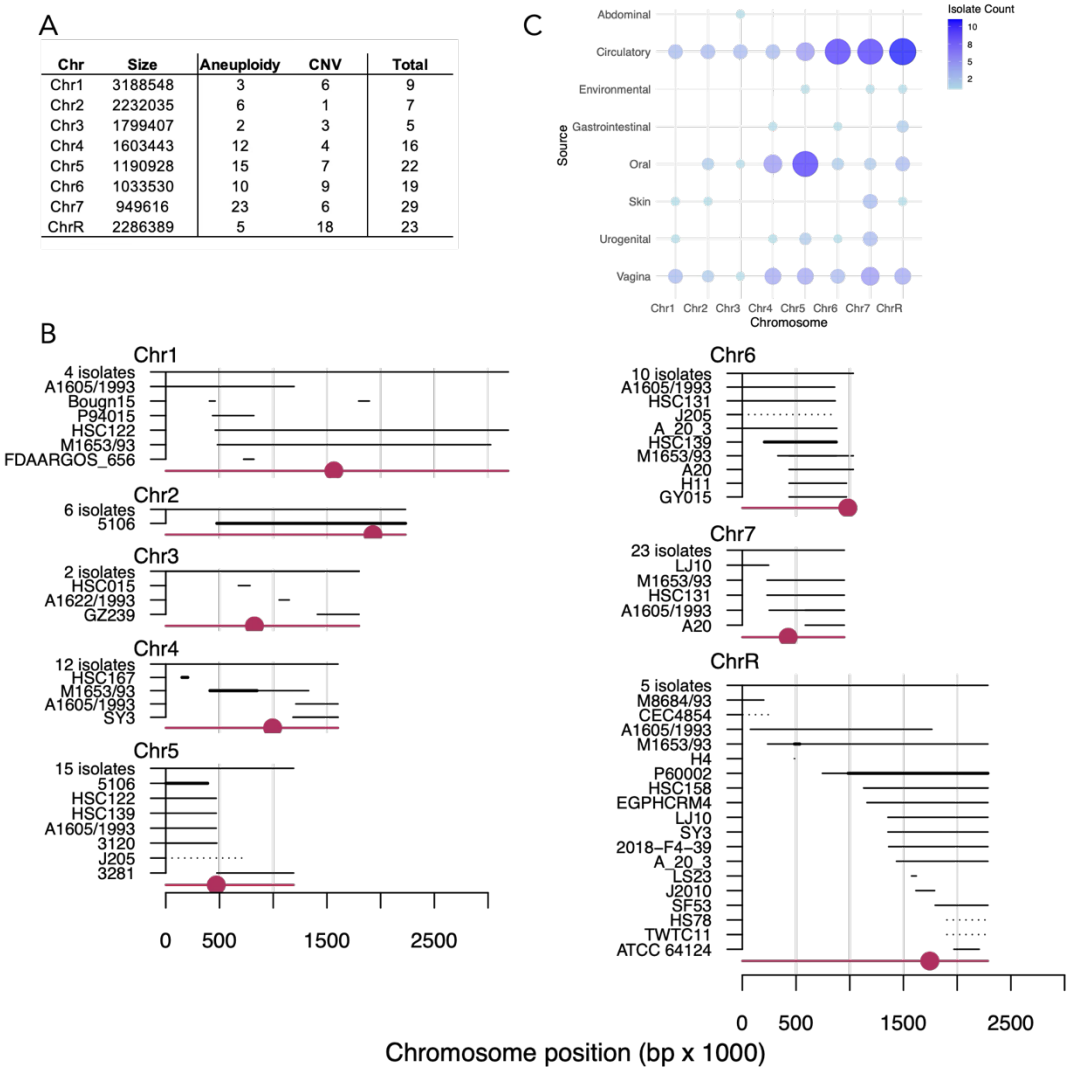


Figure 2.5: Karyotypic variation among isolates. (A) Distribution of aneuploidies and copy number variations (CNVs) across chromosomes. Karyotypic variation is more frequent in smaller chromosomes, suggesting that structural alterations may be better tolerated in shorter chromosomal regions. (B) Observed aneuploidies and CNVs in individual isolates, organized by chromosome. The length of each line represents the genomic span of the event, while the thickness indicates the magnitude of the duplication or deletion. (C) Bubble plot showing the distribution of aneuploid isolate counts across chromosomes and isolation sources. Each bubble's size and color represent the number of isolates detected for a specific chromosome-source combination. Larger and darker bubbles indicate higher counts.

2.4.7 MTL locus analyses

The distribution and variation of mating-type loci (MTL) was examined in the phylogenetically informative set of isolates (Table S1, Figure S2.8). Read alignments to the A22 reference genome were used to assess coverage across the MTLa and MTLα regions, with regions showing coverage below 0.5 interpreted as inactive or absent. As expected, the predominant genotype was the heterozygous diploid a/α, observed in 850/908 isolates (93.6%). Other genotypes were rare but present, including a/a (1.9%), α/α (3.5%), and triploid MTL configurations in chromosome 5 aneuploids (a/a/a: 0.1%; a/a/α: 0.4; a/α/α: 0.3%; α/α/α: 0.1%). Homozygous MTL isolate are from every continent and all major isolation sources. A statistically significant but modest associations were observed between MTL genotype and continent (Fisher's Exact Test with Monte Carlo simulation, $p = 0.016$) and MTL genotype and clade ($p = 0.012$), while there was no association between MTL genotype and isolation source ($p = 0.075$). That isolation source is not significant, suggests that neutral processes likely drive variation in the distribution of MTL genotypes.

2.4.8 Heterozygosity analyses

Heterozygosity was assessed across the 744 isolates with genomic information for ≥ 85% of sites. The mean genome-wide heterozygosity was 0.0065 ± 0.001 (SD) and ranged from 0.0029 to 0.014. To look at the potential effect of different factors on genome-wide heterozygosity, we did a three-factor ANOVA with isolation source, clade, and geography. There was a significant interaction between isolation source and clade in genome-wide heterozygosity across all isolation sources, while geography was not significant as either a main effect or in an interaction term. We thus dropped geography and re-ran a two-factor

ANOVA, finding that both factors were significant as both main effects and their interaction (Two-factor ANOVA test; isolation source: $F_{9, 615} = 17.4$, $p < 0.0001$, clade: $F_{20, 615} = 14.5$, $p < 0.0001$, isolation source × clade: $F_{85, 615} = 1.7$, $p < 0.0001$, Figure 2.6A). A post-hoc comparison identified that circulatory isolates had significantly higher heterozygosity than isolates from many other sites, though the isolate distributions across different sites were very overlapping (see Figure 2.6A for statistical results). The analyses of the heterozygosity ratio across the eight chromosomes revealed variability in the mean values, with chromosome 6 exhibiting the highest average (0.00814 ± 0.00324), while chromosome R had the lowest (0.00572 ± 0.00152) (Table S2.2).

Although the majority of isolates have a relatively similar level and normally-distributed level of genome-wide heterozygosity, there were several outlier isolates that had significantly lower and higher heterozygosity (Figure 2.6B). To determine whether specific regions of the genome were causative, we divided the genome into 5kb bins and counted the number of heterozygous positions in each for each isolate. There were many small LOH events across the isolates (Figure S2.10). The LOH regions encompassed events which occurred before clade expansion and hence are in common to all isolates (although the highly heterozygous isolates have heterozygous SNPs in this region Figure 2.6D), later events that are common to only some members of the clade, and isolate-specific events. Notably, the terminal end of chromosome R was consistently depleted for heterozygosity across most isolates, indicating this was its an ancestral LOH event prior to the most recent common ancestor. To see whether there were major clade-specific LOH events, for each bin we averaged across all isolates within each clade (Figure 2.6C). There were small regions of reduced heterozygosity across the genome common among isolates of the same clade

(Figure 2.6C, Figure S2.10). The seven high outlier isolates had clearly higher heterozygosity throughout their genomes compared to the clade averages (Figure 2.6D). These isolates are all from China except for one from the US and are predominantly circulatory (though one is oral, and one is gastrointestinal).

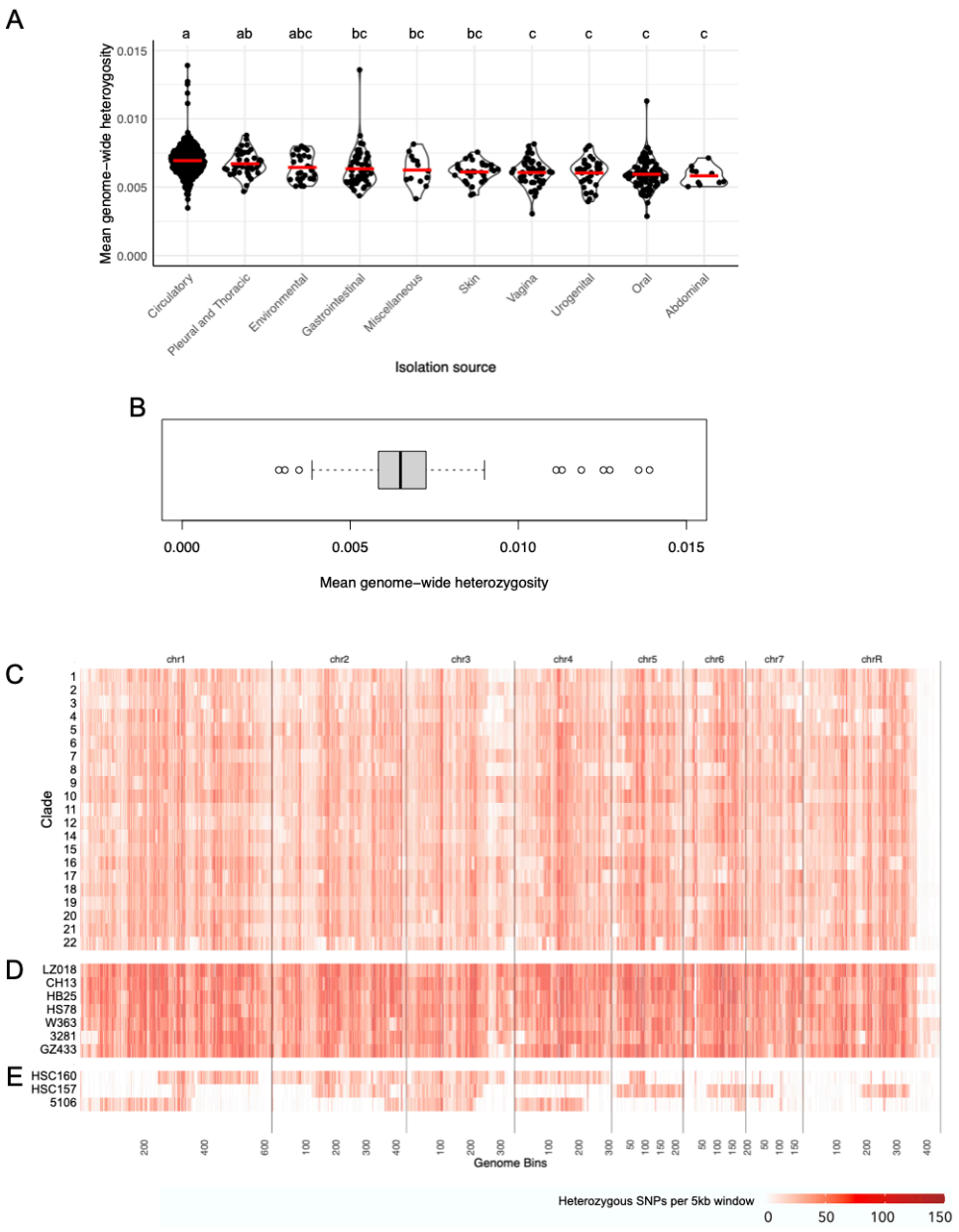


Figure 2.6: Genome-wide heterozygosity among isolates.
(A) Average genome-wide heterozygosity varies across different isolation sources. Letters above each violin plot indicate statistically significant differences based on a post hoc Tukey's test following a two-factor ANOVA; groups that do not share a letter differ significantly. B) Boxplot showing the distribution of average genome-wide heterozygosity across all isolates. (C-E) Density plots of the number of heterozygous SNPs in 5 kb windows across the genome. Each row represents an isolate, and vertical black lines indicate chromosome boundaries (chromosomes 1-7 and R). The scale bar reflects the density of heterozygous SNPs per 5 kb window, from none (white) to high density (dark red). C) Average heterozygosity from all isolates in each clade in each bin. D) Outlier isolates with high average genome-wide heterozygosity. E) Outlier isolates with low heterozygosity, largely due to large LOH regions.

By contrast, the three low heterozygosity isolates are from three different sources (oral, vaginal, circulatory) from North America. Surprisingly given the low sample size, they exhibit quite different patterns across the genome, with nearly all chromosome arms (except chromosome 3) represented in reduced heterozygosity regions.

To look for regions of both high and low heterozygosity across all isolates, we examined the average heterozygosity within each 5 kb bin across all isolates and examined the resulting histogram of the regions. Based on this, we defined regions of low heterozygosity, defined as < 5 SNPs per window. This identified 90 bins; the majority (n = 78) are at the terminal right end of chromosome R. Four additional regions are located on chromosomes 1 (one bin), 3 (nine bins), 4 (one bin) and 5 (one bin, Figure 2.7). Many named genes with known or predicted functions that could be the target of purifying selection are present (Table S2.3 at https://github.com/microstatslab/Calbians_phylogenetics/supporting_tables). There were overall fewer bins with elevated heterozygosity; defined as those as defined as > 100 SNPs per window. A total of 13 distinct genomic regions distributed across seven chromosomes (all except chromosome 3, Figure 2.7) were identified with high heterozygosity. After excluding bins that include annotated repeat sequences, we identified 26 unique protein-coding genes within these variable regions. The genes included *RAM1*, *FGR28*, *CAN1*, *RIM9*, *KAR5*, and *PEP3*, along with several proteins that have unknown functions (Table S2.3). Further work is required to validate whether there is a fitness benefit to increased heterozygosity or disruptive selection (e.g., favoring variation among isolates in different ecological contexts) in these genes.

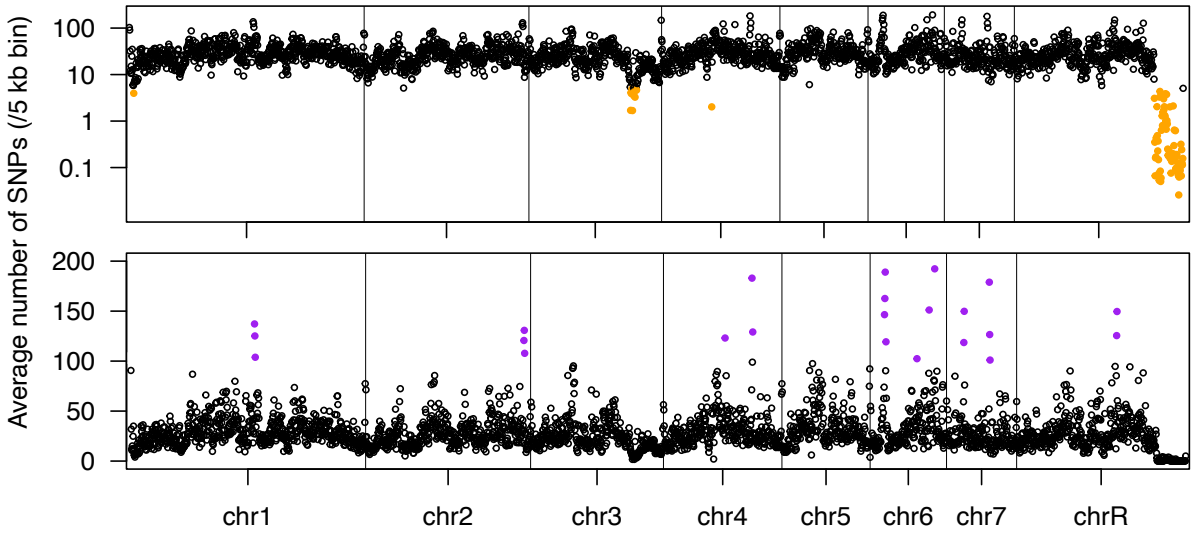


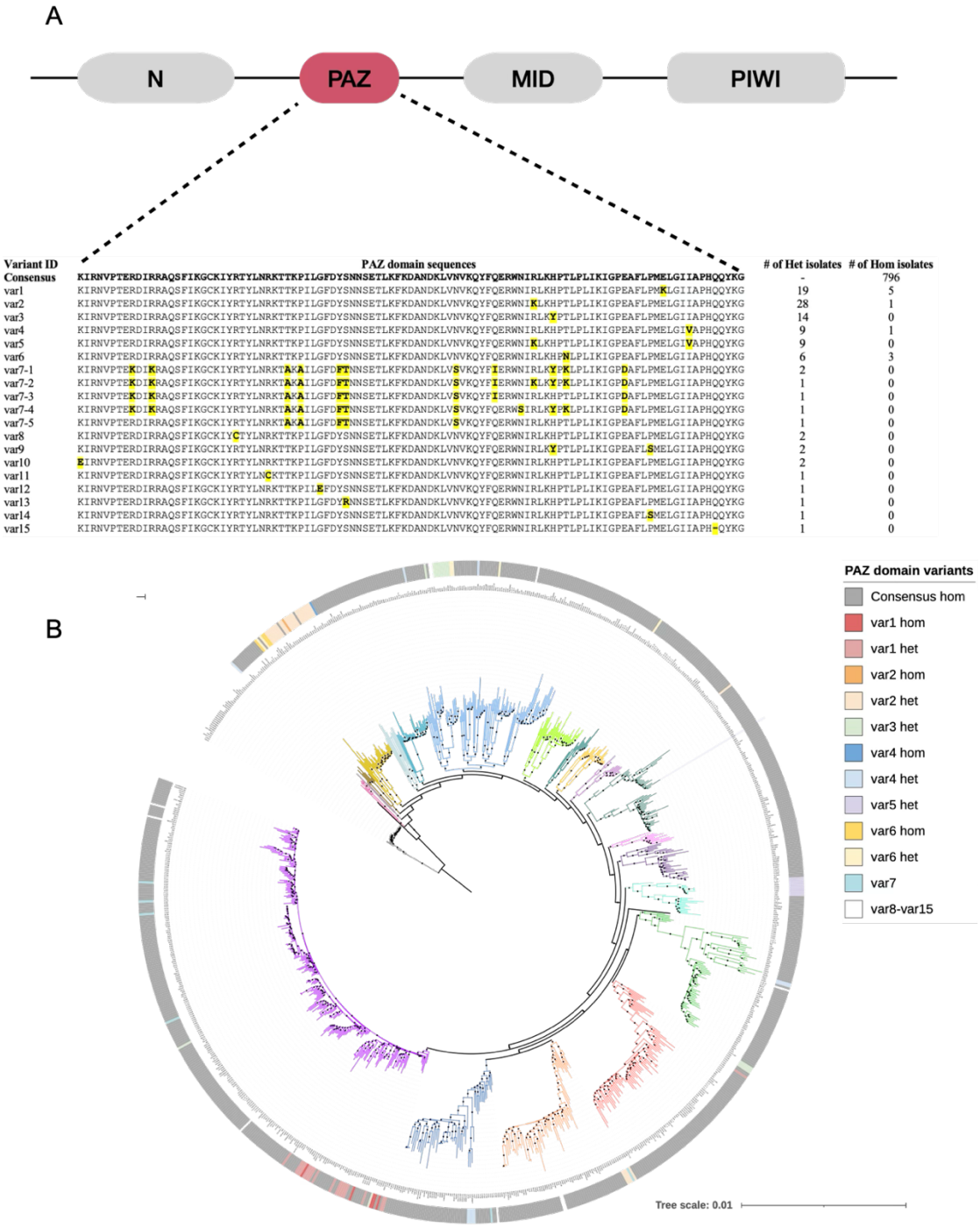
Figure 2.7: Genome-wide distribution of average number of heterozygous SNPs across 745 *C. albicans* isolates. Low-heterozygosity regions (defined as < 5 SNPs per window) are shown as orange circles, while high-heterozygosity regions (defined as > 100 SNPs per window) are indicated by purple circles. Panel A highlights the locations of high-heterozygosity regions, while panel B shows the distribution of low-heterozygosity regions across chromosomes.

2.4.9 *Ago1* PAZ domain analyses

We aimed to determine how widespread the RNAi-deficient phenotype observed in the SC5314 reference strain is among other *C. albicans* isolates, given that RNAi function is present in most strains but appears to be lost in SC5314. Among the three canonical domains conserved in Ago proteins (PAZ, MID, and PIWI), we focused on the PAZ domain, which was the one previously shown to be silenced in the SC5314 reference strain. We identified 43 unique SNPs in the PAZ domain among the phylogenetically informative set of 908 *C. albicans* (excluding the clade 13 isolates): 17 were synonymous mutations, while 26 were non-synonymous (Figure 2.8).

The RNAi-active consensus PAZ sequence was homozygous in 796 isolates (88%) of the which were distributed across the phylogenetic tree (Figure 2.8A and B). Nineteen (19) distinct PAZ domain variants (sequences) containing between 1-2 amino acid changes were identified. We identified three additional homozygous variants in the PAZ domain, in addition to the previously described var1 (K361; n = 5): var2 (K341; n = 1), var4 (N346; n = 1), and var6 (V365; n = 3). Only these four variants (var1, var2, var4, and var6) were observed in a homozygous state, indicating that the wild-type PAZ domain sequence was retained in the remaining 898 isolates.

Var1, the variant found homozygous in SC5314, was restricted to five additional clade 1 isolates, while a heterozygous version of this allele was found in an additional 19 clade 1 isolates and one clade 4 isolate (Figure 2.8B). The var2 homozygous variant was found in a single clade 20 isolate, with the heterozygous form most commonly found in the same clade. Other variants, although heterozygous, showed clade-specific enrichment: var3 was predominant in a subcluster of clade 15, while var5 was mainly linked to clade 9.



Variant ID	PAZ domain sequences	# of Het isolates	# of Hom isolates
Consensus	**KIRNVPTERDIRRAQSFIKGCKIYRTYLNRKTTKPILGFDYSNNSETLKFKDANDKLVNVKQYFQERWNIRLKHPTLPLIKIGPEAFLPMELGIIAPHQQYKG**	-	796
var1	KIRNVPTERDIRRAQSFIKGCKIYRTYLNRKTTKPILGFDYSNNSETLKFKDANDKLVNVKQYFQERWNIRLKHPTLPLIKIGPEAFLPMKLGIIAPHQQYKG	19	5
var2	KIRNVPTERDIRRAQSFIKGCKIYRTYLNRKTTKPILGFDYSNNSETLKFKDANDKLVNVKQYFQERWNIKLKHPTLPLIKIGPEAFLPMELGIIAPHQQYKG	28	1
var3	KIRNVPTERDIRRAQSFIKGCKIYRTYLNRKTTKPILGFDYSNNSETLKFKDANDKLVNVKQYFQERWNIRLKYPTLPLIKIGPEAFLPMELGIIAPHQQYKG	14	0
var4	KIRNVPTERDIRRAQSFIKGCKIYRTYLNRKTTKPILGFDYSNNSETLKFKDANDKLVNVKQYFQERWNIRLKHPTLPLIKIGPEAFLPMELGIVAPHQQYKG	9	1
var5	KIRNVPTERDIRRAQSFIKGCKIYRTYLNRKTTKPILGFDYSNNSETLKFKDANDKLVNVKQYFQERWNIKLKHPTLPLIKIGPEAFLPMELGIVAPHQQYKG	9	0
var6	KIRNVPTERDIRRAQSFIKGCKIYRTYLNRKTTKPILGFDYSNNSETLKFKDANDKLVNVKQYFQERWNIRLKHPNLPLIKIGPEAFLPMELGIIAPHQQYKG	6	3
var7-1	KIRNVPTEKDIKRAQSFIKGCKIYRTYLNRKTAKAILGFDFTNNSETLKFKDANDKLVSVKQYFIERWNIRLKYPKLPLIKIGPDAFLPMELGIIAPHQQYKG	2	0
var7-2	KIRNVPTEKDIKRAQSFIKGCKIYRTYLNRKTAKAILGFDFTNNSETLKFKDANDKLVSVKQYFIERWNIKLKYPKLPLIKIGPDAFLPMELGIIAPHQQYKG	1	0
var7-3	KIRNVPTEKDIKRAQSFIKGCKIYRTYLNRKTAKAILGFDFTNNSETLKFKDANDKLVSVKQYFIERWNIRLKHPTLPLIKIGPDAFLPMELGIIAPHQQYKG	1	0
var7-4	KIRNVPTEKDIKRAQSFIKGCKIYRTYLNRKTAKAILGFDFTNNSETLKFKDANDKLVSVKQYFQERWSIRLKYPKLPLIKIGPDAFLPMELGIIAPHQQYKG	1	0
var7-5	KIRNVPTERDIRRAQSFIKGCKIYRTYLNRKTAKAILGFDFTNNSETLKFKDANDKLVSVKQYFQERWNIRLKHPTLPLIKIGPEAFLPMELGIIAPHQQYKG	1	0
var8	KIRNVPTERDIRRAQSFIKGCKIYCTYLNRKTTKPILGFDYSNNSETLKFKDANDKLVNVKQYFQERWNIRLKHPTLPLIKIGPEAFLPMELGIIAPHQQYKG	2	0
var9	KIRNVPTERDIRRAQSFIKGCKIYRTYLNRKTTKPILGFDYSNNSETLKFKDANDKLVNVKQYFQERWNIRLKYPTLPLIKIGPEAFLSMELGIIAPHQQYKG	2	0
var10	EIRNVPTERDIRRAQSFIKGCKIYRTYLNRKTTKPILGFDYSNNSETLKFKDANDKLVNVKQYFQERWNIRLKHPTLPLIKIGPEAFLPMELGIIAPHQQYKG	2	0
var11	KIRNVPTERDIRRAQSFIKGCKIYRTYLNCKTTKPILGFDYSNNSETLKFKDANDKLVNVKQYFQERWNIRLKHPTLPLIKIGPEAFLPMELGIIAPHQQYKG	1	0
var12	KIRNVPTERDIRRAQSFIKGCKIYRTYLNRKTTKPILEFDYSNNSETLKFKDANDKLVNVKQYFQERWNIRLKHPTLPLIKIGPEAFLPMELGIIAPHQQYKG	1	0
var13	KIRNVPTERDIRRAQSFIKGCKIYRTYLNRKTTKPILGFDYRNNSETLKFKDANDKLVNVKQYFQERWNIRLKHPTLPLIKIGPEAFLPMELGIIAPHQQYKG	1	0
var14	KIRNVPTERDIRRAQSFIKGCKIYRTYLNRKTTKPILGFDYSNNSETLKFKDANDKLVNVKQYFQERWNIRLKHPTLPLIKIGPEAFLSMELGIIAPHQQYKG	1	0
var15	KIRNVPTERDIRRAQSFIKGCKIYRTYLNRKTTKPILGFDYSNNSETLKFKDANDKLVNVKQYFQERWNIRLKHPTLPLIKIGPEAFLPMELGIIAPH-QYKG	1	0

Figure 2.8: *Candida albicans* Ago1 PAZ domain sequence variants. A) Schematic representation of *Candida albicans Ago1* gene and alignment of unique PAZ domain sequences (aa 271-373) identified from 907 *Candida albicans* isolates with the consensus (known active) sequence in bold. Positions that differ from the consensus domain are highlighted in yellow. B) The distribution of the variant in the phylogeny. The variants are indicated as with homozygous (hom). or heterozygous (het).The branches of the phylogeny are colored by clade.

2.4.10 Intra-individual analysis

Nine studies included more than one isolate collected from the same individual, or from a healthy mother-infant dyads (Table S2.4). Nearly all individuals are from the United States (75) and Canada (12), four are from Spain, and one each from Morocco, France, Brazil and Tunisia. The largest study comprised oral isolates of 59 mother-infant dyads from the United States. The other studies took multiple isolates at a single time point at the same body site (12 individuals, blood, lung and oral isolates), a single time point at multiple sites (7 individuals, oral and fecal or rectal and vaginal) or multiple time points from the same or different body sites (8 individuals with oral samples, one individual with urine samples, and one individual with blood and pleural samples). The number of isolates per individual ranged from 2 to 23 (mean 3.8, median 2). Most were colonized by genetically monophyletic *C. albicans* populations, though fourteen individuals carried isolates from two clades: ten from healthy mother-infant dyads, two from sequential oral isolates taken from HIV patients, one from sequential isolates taken from urine, and two from oral and fecal samples taken from healthy individuals. One individual contained isolates from three clades (oral and fecal isolates from a healthy individual). Neither sampling strategy nor the study context influenced variation in mean genome-wide heterozygosity among isolates from the same individual (Figure S2.10), nor the likelihood of finding variation among isolates for mating locus or aneuploidy (Table S2.1).

2.4.11 Environmental isolates

The environmental isolates were collected from a number of sources, including food spoilage, tree bark, starling (bird) feces, and soil. Most environmental isolates clustered closely with clinical isolates throughout the phylogeny (Figure 2.4). One soil isolate, however, was a singleton that did not cluster with any other isolates and was the most terminal isolate on the phylogeny beside the outgroup. The two European isolates from starlings were also located very close to the edge of the phylogeny in a very small clade (clade 16), with only three other isolates (a urogenital isolate from Europe, an oral isolate from the UK and circulatory isolate from China).

The average heterozygosity of environmental isolates did not differ significantly from other sources (Figure 2.6A), nor were they more likely to be aneuploidy, MTL homozygous, or Ago1 variant. Given that 19 of the environmental isolates are from food spoilage, the true environmental representation in the phylogeny is very small. Given that three of the 'true' environmental isolates are found at the end of the current species phylogeny, it demonstrates that additional effort to sampling soil, trees, and non-human hosts is needed to better capture the true breadth and diversity of environmental isolates.

2.5 Discussion

The distribution of *C. albicans* isolates reveals significant sampling biases in both geography and isolation source. There is a notable absence of large-scale, region-specific studies that capture isolate diversity from Africa, the Middle East, and Australia, and the few isolates that do exist are only from a small number of countries. In addition to the paucity of short-read WGS data, of the ~6000 isolates submitted to the pubMLST database (as of July 30, 2025) (Sayers et al., 2022), there are only 197 isolates from Africa and 88 from Oceania. The lack of genomic studies from these locations highlights a critical gap in capturing local *C. albicans* diversity and negatively impacts our ability to assess regional genetic diversity and population structure. The types of studies that have generated WGS data are also biased, often focused on a specific disease presentations. For example, a large proportion of isolates come from studies on candidemia (particularly from Asia) (Chew et al., 2023; Gong et al., 2023), oral candidiasis (North America) (Alkhars et al., 2024; Ford et al., 2015) and commensal colonization (Alkhars et al., 2024; Anderson et al., 2023; Sitterlé et al., 2019) (predominantly oral isolates, the majority from North America). Vaginal isolates are the most evenly sampled across continents, consistent this being the most common infection type. Environmental isolates are rare and only a handful of isolates have been collected from genuinely natural environments such as soil (Opulente et al., 2019), sequences included here) or tree bark (Bensasson et al., 2019). Despite their scarcity, the environmental isolates generally show high genomic similarity to human-associated isolates. The view that *C. albicans* is transmitted strictly through vertical (mother-to-child) transmission (Caramalac et al., 2007; Filippidi et al., 2014) is somewhat challenged by finding micro-geographic association within the phylogeny. Isolates from the same place from different isolation

sources often cluster together, suggesting the potential for alternative routes of transmission. One study found that the majority of mother-child dyads in early life tend to have very closely related isolates (Alkhars et al., 2024) (this is also supported by our phylogenetic analysis, Table S2.1), it would be interesting to look at isolates over a longer period of time, to determine their stability. There is a great need for expanded environmental sampling to better understand the factors that drive relatedness in the *C. albicans* phylogeny and to tease apart the currently confounding influences of geography and isolation source.

We utilized phased whole-genome data paired with a statistical tool to construct a phylogeny and identify 20 clades (with only two singletons). The phylogenetic structure is largely consistent with the previous WGS tree (Ropars et al., 2018), yet identified six additional clades. The isolates in new clades are predominantly from Asia. A recent study added 369 clinical isolates from China, proposed 38 clades, including 21 novel clades (Gong et al., 2023). In our analyses, the novel clades described by Gong *et al.* were either merged into previously existing clades or part of the six novel clades we described. Given the historical legacy of MLST clade names left five numbers unaccounted from in WGS phylogenies, we propose to reassign these "retired" MLST clade numbers to the novel clades identified. We also proposed removing the alphabetic designations from (Ropars et al., 2018) to numeric names to fully standardize clade nomenclature, minimize confusion, and conservatively maintain continuity with prior studies.

The lack of strong geographic or isolation site-specific patterns in chromosomal aneuploidy and copy number variations (CNVs) suggests that these genomic alterations likely arise from local or individual-level selective pressures, rather than being driven by broader population structure or regional factors. We found that 9.47% of isolates had a

major karyotypic variation, very similar to the frequency of the Ropars et. al strain set (Ropars et al., 2018). The specific chromosomes involved differed between circulatory and oral isolates (chromosomes 6, 7, and R for circulatory; chromosomes 4 and 5 for oral). Trisomy of chromosome 5 is thought to be selected for during oropharyngeal candidiasis, as it facilitates a commensal-like phenotype (Forche et al., 2019). However, studies have also shown that chromosome 7 trisomy can enhance colonization of both the gastrointestinal tract and the oral cavity in mouse models (Ene et al., 2018; Kakade et al., 2023; Mishra et al., 2025). It may also be that some or many of the observed karyotypic variants are neutral or depend heavily on the genetic background. Karyotypic variants were observed throughout the phylogeny and were unclustered, supporting the idea that aneuploidy and CNVs are transient. It is also likely that their effects (if there are any) are strain-dependent. For example, although trisomy of chromosome 4 was shown to confer fluconazole resistance in a clinical *C. albicans* isolate (Anderson et al., 2017), aneuploidy was also observed in isolate T118 during serial passage in fluconazole, yet did not directly contribute to enhanced drug resistance (Selmecki et al., 2009).

The distribution of mating-type locus (MTL) genotype was also relatively consistent with previous large-scale surveys. The α/α and a/a genotypes were relatively rare (3.5% and 1.9%, respectively), compared to previously reported homozygosity rates of 2.2% (Lockhart et al., 2002; Ropars et al., 2018)) and 3.2% (Lockhart et al., 2002; Ropars et al., 2018) (though Odds et al. (2007) observed a slightly higher prevalence of 8.5%). In all studies, including this one, the α/α genotype was more common than a/a, suggesting that the loss of the MTLα locus may occur more frequently or be more tolerable in the *C. albicans* population.

The average genome-wide heterozygosity among our isolates was 0.0065 (6.5 heterozygous SNPs per 1000 bases). This was nearly identical to the average of three isolates from oak trees (0.0066, (Bensasson et al., 2019)), and an analysis of 61 diverse isolates (0.0067, (Mixão and Gabaldón, 2020)) while higher than the average of 182 isolates from (0.0048, Ropars et al. 2018)). This is likely due to differences in methodology, rather than strain-set.

We found heterozygosity varied across anatomical sources, indicating that host environments may influence genetic diversity. Circulatory isolates, for example, displayed higher heterozygosity than gastrointestinal or oral isolates, potentially reflecting the immune-challenged and dynamic nature of the circulatory environment. Lower heterozygosity in isolates from localized infections, such as abdominal or gastrointestinal sources, suggests more stable pathogen populations in these sites. LOH analyses revealed both ancestral and clade-specific events, including a conserved LOH region on chromosome R and clade-restricted LOH on chromosome 3 (clades 1, 4, 8). Differences in heterozygosity among *C. albicans* clades have previously been reported (Gong et al., 2023; Hirakawa et al., 2015; Ropars et al., 2018), yet these differences are potentially driven by biased sampling, with the confounding variable of isolation source nested into clade.

The results from this study reveal different genetic alterations in the PAZ domain of *Candida albicans* Ago1. A very specific mutation (Ago1-K361) in SC5314 and 7 other isolates has been shown to result in RNA interference (RNAi) deficiency in this pathobiont (Iracane et al., 2024). The identification of 43 SNPs, including 17 synonymous and 26 non-synonymous mutations, underscores the genetic variation within this domain. The majority of isolates (> 99 %) maintained the RNAi-active consensus PAZ sequence, with 88% of

isolates being homozygous for this sequence, indicating that RNAi activity is preserved in most *Candida albicans* isolates. However, the study also highlights the presence of multiple PAZ domain variants which carry mutations that deviate from conserved residues, potentially impairing Ago1 function. All but four of the variants identified here were heterozygous with the active consensus sequence. Thus, such variants are potentially pseudogenes, allowing for the accumulation of mutations while the active form remains functional. Previous studies identified the Ago1-K361 variant (var1) as responsible for RNAi deficiency in the SC5314 reference strain (Iracane et al., 2024), a mutation that also abolishes RNAi activity in *Caenorhabditis elegans* (Tabara et al., 1999). This study extends these findings by demonstrating the presence of several additional PAZ domain variants. The RNAi-defective var1 variant, found in 2.7% of isolates, was previously reported in 3% of isolates (9/296), all confined to clade 1 (Iracane et al., 2024). The lack of clear clustering of these isolates in the phylogeny suggests that RNAi-deficient variants may arise independently through convergent evolutionary pathways; however, their limited spread across the population also implies a potential fitness cost or selective constraint associated with loss of RNAi function. Interestingly, while var1 mutants exhibited increased TLO gene expression (which is linked to enhanced resistance to oxidative stress (Flanagan et al., 2018)), their fitness under standard laboratory conditions was unaffected, suggesting that the functional relevance of Ago1 inactivation may be more pronounced in natural or *in vivo* environments rather than in controlled lab settings. In this study, we identified three additional homozygous variants var2 (K341), var4 (N346), and var6 (V365). A heterozygous form of var6 has previously been reported in six isolates (Iracane et al., 2024). As this study was mainly *in silico,* we did not assess the biological impact of these variants. In the future, it

will be important to explore the differential functional impacts of the various variants to better understand their roles in *C. albicans* biology.

Our strain set included 95 individuals with more than one isolate. However, the majority of individuals contributed fewer than six isolates (the recommended number to explore variation (Adamu Bukari et al., 2025)), limiting our ability to robustly characterize within-host population diversity in most cases. Nevertheless, the observed clade diversity in a subset highlights the potential for intra-host microevolution or mixed-strain persistence, with implications for treatment and resistance surveillance, which could potentially complicate treatment and resistance monitoring.

2.6 Conclusion

This study refines the phylogenetic structure of *C. albicans* using phased whole-genome data, resolving both known and novel clades with greater clarity and proposing standardized definitions to improve cross-study comparisons. While clade distribution appears to show geographic structure, our results suggest this pattern is largely confounded by differences in isolation source. The predominance of the a/α MTL configuration supports largely clonal reproduction, while the presence of MTL homozygotes indicates a potential for recombination. Aneuploidy and CNVs lack consistent geographic or ecological patterns, suggesting they may reflect transient responses to environmental stress. Although environmental isolates are underrepresented, their genetic overlap with human-associated strains and broader clade diversity (particularly in Europe) point to the environment as a potential reservoir and challenge models of strictly vertical transmission. Lastly, the discovery of novel mutations in RNAi components, including Ago1, highlights the need for

functional studies to understand their impact on gene regulation and host interactions. Together, these findings provide a framework for *C. albicans* phylogenetics for exploring the evolutionary and ecological dynamics of *C. albicans*.

Chapter 3: Global whole-genome phylogenomics of *Nakaseomyces glabratus* reveals admixture and refines sequence type-based classification

This chapter will be submitted to a peer-reviewed journal article:
Contributing authors: Adamu Bukari, A. R., Brooke, and Gerstein, A. C.

3.1 Abstract

Nakaseomyces glabratus is a globally distributed opportunistic yeast. An ongoing discussion in *N. glabratus* population structure studies has been whether genetic clusters should be defined using multilocus sequence typing (MLST) or whole-genome sequencing (WGS). To assess the concordance between MLST-based and WGS-based phylogenetics, we analyzed a dataset of 548 *N. glabratus* whole genome sequences acquired from 12 countries. We also determined the prevalence of admixture, aneuploidies and copy number variations within the isolate set. We found that WGS-defined clusters largely recapitulate the MLST topology. Fourteen clusters are comprised of a single MLST sequence type (STs) and the remaining clusters contain STs with very closely related allele profiles. Based on this, we propose a pragmatic naming convention that retains cluster labels based on the primary ST, consistent with the system used in other microbial species. We identified 65 admixed isolates with multiple ancestries: 7 are singletons, while 58 are from six different clusters. The admixed isolates were geographically widespread and found predominantly within two clusters, revealing a higher level of gene flow across clusters than has previously been appreciated. We additionally detected aneuploidy in 4% of isolates, most commonly in chrE, which contains ERG11, the gene encoding the enzyme targeted by azole antifungals. Copy number variants, some of which co-occurred with aneuploidies, were primarily identified on chrD, chrE, chrI, and chrM. Our findings underscore the utility of WGS for high-resolution population structure analyses and demonstrate that deep splits between clusters explains the utility of MLST designations. More balanced global sampling will be critical to fully understand the diversity and evolution of *N. glabratus*.

3.1 Introduction

Nakaseomyces glabratus is an opportunistic fungal pathogen of increasing clinical importance, particularly in the context of healthcare-associated infections among immunocompromised individuals. In recognition of its growing threat, the World Health Organization classified *N. glabratus* in 2022 as a high-priority fungal pathogen (fifth overall, (WHO, 2022)). Epidemiological trends over the past two decades indicate an increasing prevalence of *N. glabratus* as the second most common invasive yeast. infection after *Candida albicans*, particularly in North America, Australia, Europe and the Middle East (Arastehfar et al., 2021; Astvad et al., 2018; Chapman et al., 2017; Fuller et al., 2019; Kord et al., 2020; Pfaller et al., 2009, 2019; Taj-Aldeen et al., 2014). In comparison to *C. albicans*, *N. glabratus* isolates exhibit innate resistance to azoles more frequently (Oxman et al., 2010), and they can also evolve resistance to echinocandins (Pfaller et al., 2012), the internationally recommended frontline antifungal for *N. glabratus* (Pappas et al., 2009). *N. glabratus* infections can thus be challenging to treat, which potentially explains why they continue to increase in incidence.

N. glabratus belongs to the Nakaseomyces clade (Fitzpatrick et al., 2006; Gabaldón et al., 2016; Kurtzman and Robnett, 2003), a group that is more closely related to the baker's yeast *Saccharomyces cerevisiae* than to *C. albicans*. It is a haploid yeast that has widely been considered asexual, as all attempts to observe *N. glabratus* mating in the laboratory have so far been unsuccessful. However, there may be a cryptic sexual cycle (Gabaldón and Fairhead, 2019) as there are distinct mating types (Boisnard et al., 2015; Brisse et al., 2009; Carreté et al., 2018; Dodgson et al., 2005; Fabre et al., 2005; Kaplan et al., 2019; Muller et al., 2008; Srikantha et al., 2003), mate-type switching has been observed (Boisnard et al., 2015; Brisse et al., 2009; Carreté et al., 2018; Dodgson et al., 2005; Fabre et al., 2005; Kaplan et al., 2019;

Muller et al., 2008; Srikantha et al., 2003), and admixture analysis (in both MLST and WGS data) is consistent with the presence of sexual recombination (Boisnard et al., 2015; Brisse et al., 2009; Carreté et al., 2018; Dodgson et al., 2005; Fabre et al., 2005; Helmstetter et al., 2022; Kaplan et al., 2019; Muller et al., 2008; Srikantha et al., 2003).

Genetic diversity among *N. glabratus* isolates has been studied by various techniques, including polymorphic locus sequence typing, pulsed field gel electrophoresis (Bennett et al., 2004; Lin et al., 2007), amplified fragment length polymorphism analysis (Paluchowska et al., 2014), multilocus microsatellite typing (Abbes et al., 2012) and multilocus sequence typing (MLST) (Dodgson et al., 2003; Lott et al., 2010). Most common among these techniques is MLST using the coding regions of six loci (*FKS*, *LEU2*, *NMT1*, *TRP1*, *UGP1*, and *URA3*). These six loci are conserved housekeeping genes that were selected because they showed the greatest variation among isolates and, when combined, produced the highest number of distinct sequence types (Dodgson et al. 2003). *FKS* encodes a subunit of β-1,3-glucan synthase, involved in fungal cell wall biosynthesis; *LEU2* and *TRP1* participate in amino acid biosynthesis (leucine and tryptophan, respectively); *NMT1* encodes N-myristoyltransferase, essential for protein lipid modification; *UGP1* encodes UDP-glucose pyrophosphorylase, a key enzyme in carbohydrate metabolism; and *URA3* encodes orotidine 5′-phosphate decarboxylase involved in pyrimidine biosynthesis. These loci provide sufficient sequence diversity to discriminate among isolates while maintaining stability across the population, making them well-suited for strain-level typing in epidemiological and population genetic studies (Dodgson et al., 2005; Lott et al., 2010). As of May 2025, the PubMLST database (https://pubmlst.org), the comprehensive global repository for microbial molecular typing and genome diversity (Jolley et al., 2018), contains 2,187 *N.*

glabratus isolates from 314 sequence types (STs). The most recent MLST studies identified seven distinct clusters (Dodgson et al., 2005).

Recently, a phylogeny from whole-genome sequence (WGS) data from 34 globally sampled *N. glabratus* isolates also recovered seven genetically distinct clades (Carreté et al., 2018) Interestingly, they found a much deeper intra-clade divergence and greater between-clade genetic diversity compared to *C. albicans*. The MLST and WGS topologies partially but not completely overlapped, demonstrating the utility of WGS in providing greater precision in identifying relatedness among isolates (Carreté et al., 2018). Additionally, contrary to the MLST phylogenies, which had indicated geographical enrichments in some clades, the WGS phylogeny revealed a lack of strong geographical structure. In support of a cryptic sexual cycle, the 34-isolate set (Carreté et al., 2018), as well as an analysis of 151 WGS isolates (Helmstetter et al., 2022), both uncovered evidence of population admixture between clusters.

A recent study by Zheng et al. (2022) used flow cytometry to analyze the ploidy of 500 clinical isolates and found that approximately 4 % of them displayed aneuploid, diploid, or polyploid forms. Among these, 3 % maintained a stable diploid state, meaning that the diploid chromosomal content persisted over time in serial culture transfers in YPD medium *N. glabratus* has 13 chromosomes; previous studies that observed aneuploidies (typically disomies) found that they are predominantly in chromosomes C, E, G and J (Carreté et al., 2018, 2019). ChrJ aneuploidies were observed to arise during growth in rich, antifungal-free media (Carreté et al., 2018), yet aneuploidies are typically lost under non-selective conditions - supporting a potentially transient role in adaptation to drug resistance (Rustchenko, 2007). However, in a more recent experimental evolution study, Ksiezopolska

et al. (2024) found that fluconazole-resistant replicate lines retained ChrE aneuploidy even after extensive propagation; specifically, after 35 serial passages in non-selective YPD medium. This suggests that, although often transient, certain aneuploidies in *N. glabratus* can be maintained stably in the absence of drug pressure.

In this study, we generated a new WGS phylogeny from 548 clinical isolates from 12 countries. This larger dataset enabled us to evaluate the saturation and resolution of a WGS phylogeny compared to the MLST database. While the resulting WGS phylogeny generally reflected sequence type (ST)-based groupings, many clusters encompassed multiple STs, thereby challenging recent proposals of strictly using ST designations. We provide a reconciliatory ground for clade designation by naming clusters by their dominant STs. Furthermore, we identified 65 isolates with signatures of admixture, providing additional evidence of gene flow across genetically distinct lineages, thus adding to the evidence of a sexual cycle in *N. glabratus*.

3.2 Materials and methods

3.2.1 Isolate collection

We downloaded data from 528 paired-end Illumina fastq files from 20 BioProjects from the NCBI SRA; we refer to these isolates as the "*N. glabratus* WGS isolates". Approximately half were previously published in 16 articles (Barber et al., 2019; Biswas et al., 2018; Bukari et al., 2023; Carreté et al., 2018, 2019; Galocha et al., 2022; Guo et al., 2020; Håvelsrud and Gaustad, 2017; Helmstetter et al., 2022; McTaggart et al., 2020; Pais et al., 2022; Salazar et al., 2022; Siscar-Lewin et al., 2021; Szarvas et al., 2021; Vale-Silva et al., 2017; Xu et al., 2021) while 287 isolates have no publication associated with them (PRJNA596170, PRJNA593955, PRJNA329124, PRJNA524686, Table S3.1). In addition, 18 isolates from 2012 were acquired from the microbiology lab at Health Science Centre (with assistance from the Shared Health/Diagnostic Services) in Winnipeg, Canada (Table S3.1). For those isolates, genomic DNA was extracted from single colonies of each using the phenol-chloroform protocol previously described (Kukurudz et al., 2022). DNA quality and concentration were assessed spectrophotometrically (NanoDrop 2000, Thermo Scientific™) and fluorometrically (Qubit® 2.0 Fluorometer with the dsDNA BR Assay Kit, Invitrogen™), respectively. The genomes were sequenced by the Microbial Genome Sequencing Center (Pittsburgh, USA) using the Illumina NextSeq 550 sequencing technology with paired-end reads of 150 bp. The bcl-convert v3.9.3 package (https://support-docs.illumina.com/SW/BCL_Convert/Content/SW/FrontPages/BCL_Convert.htm) was used in demultiplexing, quality control, and adapter trimming. The reads have been deposited at the National Center for Biotechnology Information (NCBI) Sequence Read Archive under BioProject ID PRJNA991137. The paired-end Illumina reads from a

Nakaceomyces bracarensis (SRR25783652) isolate described by Marcet-Houben et al. (2024) was initially included as an outgroup. However, this isolate was found to cluster closely with other *N. glabratus* isolates, so it was ultimately not included.

3.2.2 *In silico* MLST typing

We assigned sequence types (STs) to the *N. glabratus* WGS isolates using stringMLST with default parameters (Gupta et al., 2017), leveraging its capability to retrieve the latest MLST allele and profile definitions directly from the PubMLST database (Jolley et al., 2018). The analysis employed the established six-locus MLST scheme for *N. glabratus* (*FKS, LEU2, NMT1, TRP1, UGP1*, and *URA3*; note that PubMLST continues to label this species as *C. glabrata*, (Denning, 2024)). A request has been submitted to PubMLST to update the species name to *Nakaseomyces glabratus* in line with current taxonomic revisions. One profile combination did not match any existing entries in the database and was submitted to the database as a new ST.

3.2.3 Variant calling

The sequence reads from all WGS isolates were trimmed with Trimmomatic (v0.39) (Bolger et al., 2014) with standard parameters (LEADING: 10, TRAILING: 3, SLIDINGWINDOW:4:15, MINLEN: 31, TOPHRED33). Quality was assessed with FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC (Ewels et al., 2016). Trimmed paired-end reads were mapped using bwa-mem (Li, 2013) to the CBS 138 reference genome (GCA000002545v2) downloaded from the Ensembl Genome Database (Yates et al., 2022). The resulting SAM file was coordinate-sorted and converted to a Binary

Alignment Map (BAM) file using samtools v1.9 (Li et al., 2009). Alignment quality was assessed with CollectAlignmentSummaryMetrics from Picard v2.26.3 (http://broadinstitute.github.io/picard) and consolidated across all samples with MultiQC (Ewels et al., 2016). All files had a > 95% mapping quality. BAM files were further processed with Picard to add a read group annotation so that samples with the same BioProject ID had the same read group, to remove duplicate PCR amplicons, and to fix mate pairs. The average coverage for each isolate was estimated using samtools v1.9 (Li et al., 2009). We also quantified chromosomal aneuploidy and sub-chromosomal copy number variation (CNV) for each WGS isolate. Sequence reads were aligned to the reference genome, and per-base coverage was calculated from the resulting BAM files using samtools depth. A custom R script (available at https://github.com/MicroStatsLab/Microstats/binCoverage.R) then partitioned the genome into non-overlapping 5 kb bins and calculated the average read depth per bin. The average coverage across each of the 13 chromosomes was calculated for each strain. For each isolate, the bin values were normalized to a haploid base ploidy by dividing by the median chromosome coverage.

Aneuploidies and regions with increased copy number (CNVs) were detected from normalized coverage profiles through visual inspection of coverage plots using a custom script (available at https://github.com/MicroStatsLab/Microstats/SWLine.R). When the normalized coverage deviated from an integer value for an entire chromosome or CNV region, the normalized data was re-examined assuming a diploid or triploid base ploidy (i.e., by multiplying the normalized data by 2 or 3). CNV breakpoints were manually determined as the bin where a distinct, abrupt shift in normalized read depth was observed.

The GATK Best Practices were adapted for variant calling. In sequence, HaplotypeCaller, CombineGVCFs, GenotypeVCFs, VariantFiltration, and SelectVariants (DePristo et al., 2011; Poplin et al., 2018; Van der Auwera et al., 2013) were used to identify single-nucleotide variants (SNPs) among all sequenced isolates in haploid mode. The resulting SNP table was hard filtered using the suggested GATK parameters (QualByDepth < 2.0, FisherStrand > 60.0, root mean square mapping quality < 30.0, MappingQualityRankSumTest < −12.5, ReadPosRankSumTest < −8.0). We excluded variants that were called in known repetitive regions of the genome, as these are likely to reflect sequencing misalignments rather than true variants, i.e., the subtelomeric regions (15kb from the start and end of each chromosome), the centromeres, and the major repeat sequence regions (Table S3.2)

To assess heterozygosity in aneuploid chromosomes, for the relevant isolates, the SNP calling analysis was repeated in diploid mode in HaplotypeCaller to determine the number of called homozygous and heterozygous SNPs after filtration. To compare the prevalence of heterozygous SNPs in aneuploid chromosomes to the number of heterozygous SNPs called due to base calling errors, we also re-ran 18 euploid haploid isolates, matched from the same studies, in diploid mode.

3.2.4 Phylogeny construction

The multi-sample VCF file, was converted to a FASTA alignment using a publicly available Python script that creates an alignment matrix for phylogenetic analysis (vcf2phylip.py v2.8, downloaded from https://github.com/edgardomortiz/vcf2phylip) (Ortiz, 2019). The FASTA alignment was parsed in FastTree (2.1.11) (Price et al., 2010) in

the double-precision mode to construct an approximate maximum-likelihood phylogenetic tree using the general time-reversible model and the -gamma option to rescale the branch lengths. The phylogeny was visualized and annotated with the Interactive Tree Of Life (iTOL, v5) (Letunic and Bork, 2021). The phylogeny was midpoint-rooted, as the expected *N. bracarensis* isolate could not serve as an appropriate outgroup due to its unexpected placement internal to the *N. glabratus* isolates.

3.2.4 TreeCluster cluster designations

Cluster designations were determined using TreeCluster (Balaban et al., 2019), as described in Chapter 2. Briefly, this tool partitions phylogenetic trees into clusters based on user-defined thresholds under eight clade-respecting strategies (avg clade, leaf dist max, length clade, max clade, med clade, root_dist, single-linkage, sum branch). Cluster predictions from four strategies were chosen for further investigation because they had the greatest stable cluster number prediction over consecutive thresholds. To compare the four strategies, we selected the threshold from each range with the least number of singleton isolates. The clustering outputs from each of the chosen strategies were then compared with the previously ST assignment from the *in silico* analysis.

3.2.5 Admixture analyses

We assessed population structure in *N. glabratus* using ADMIXTURE (Alexander et al., 2009) following the methods previously used in *N. glabratus* (Helmstetter et al., 2022). Input data were prepared by converting the variant call format (VCF) into PLINK BED format. Nonstandard contig names (i.e., chromosomes), which were initially labelled A-M following

N. glabratus standard nomenclature was changed to be sequential integers (1-13) as required for ADMIXTURE.

Admixture analysis was performed on haploid mode across a range of hypothetical numbers of ancestral populations (K values) from 1 to 40. Five-fold cross-validation was employed to evaluate model performance for each K; the cross-validation error outputs were parsed using a custom R script to determine the K value with the lowest error. The admixture population proportions for the optimal value of K were visualized to determine ancestral contributions.

3.2.5 **Data and script availability**

All data, supplementary tables, and scripts to run statistical analyses and generate figures are available at

https://github.com/microstatslab/Nglabratus_phylogenetics.

Supplementary tables can be found at

https://github.com/microstatslab/Nglabratus_phylogenetics/supporting_tables

3.3 Results

We analyzed a total of 548 *N. glabratus* isolates. The sampling is geographically biased; although isolates were collected from four continents, the majority came from North America (59.0%), followed by Europe (31.4%), Oceania (9.1%), and Asia (0.4%) (Figure 3.1). Similarly, although the isolates were obtained from a range of clinical sources, over three-quarters came from the circulatory system (79.2%). Additional sources included the gastrointestinal tract (2.6%), oral cavity (2.4%), and urogenital tract (1.8%), with less than 1% of isolates from the abdominal, pleural/thoracic, skin, vaginal, or other sites (note that anatomical source information was lacking for 53 isolates; Table S3.1).

Figure 3.1: Distribution and literature source of isolates used in this study. A) Geographical distribution of the 548 isolates analyzed in this study. Circle size is proportional to the log of the number of isolates from each region, while colored sectors represent their respective isolation sources. White sectors are isolates with unknown isolation sources. The image was generated using Microreact. B) Publication source of the isolates used the number of isolates from each manuscript is provided in brackets.

3.3.1 *N. glabratus* phylogenetic clusters

A total of 82,564 SNP positions were identified across all isolates and used to construct the phylogeny. To identify clusters, we applied eight tree-based clustering strategies in TreeCluster (Balaban et al., 2019), each across the range of possible threshold values. The number and composition of clusters varied by method and threshold space (Figure 3.2A), with the number of identified potential clusters ranging from 1 to 200. Three of the strategies identified 27 clusters, all across many consecutive threshold values (single linkage, 379 consecutive thresholds; maximum clade, 353 thresholds; length clade, 219 thresholds; Figure 2A). The next best supported number of clusters was 18, identified from the medium clade strategy across 375 consecutive thresholds. By comparison, the pubMLST database currently contains 269 STs.

The 27-cluster phylogenies varied in the number of singleton isolates, with the majority of branch length threshold space (reported as substitutions per site) identifying 9 singleton isolates (single linkage: 0.00354—0.004; max clade: 0.0099—0.0361; length clade: 0.0023—0.018), with a smaller amount of threshold space supporting 8 or 7 singletons. To evaluate the differences among phylogenies we visualized phylogenetic assignments for the minimum and maximum threshold value identifying 7, 8, or 9 singletons from each of the three strategies. The same isolates (HSC003 and WM_18.50) at the threshold extremes were repeatedly classified as either singletons or grouped into clusters (Figure S3.1). These isolates visually appear as distinct on the phylogeny. Accordingly, we use the majority-supported, 27 cluster, 9 singleton isolate tree (Figure 3.2A).

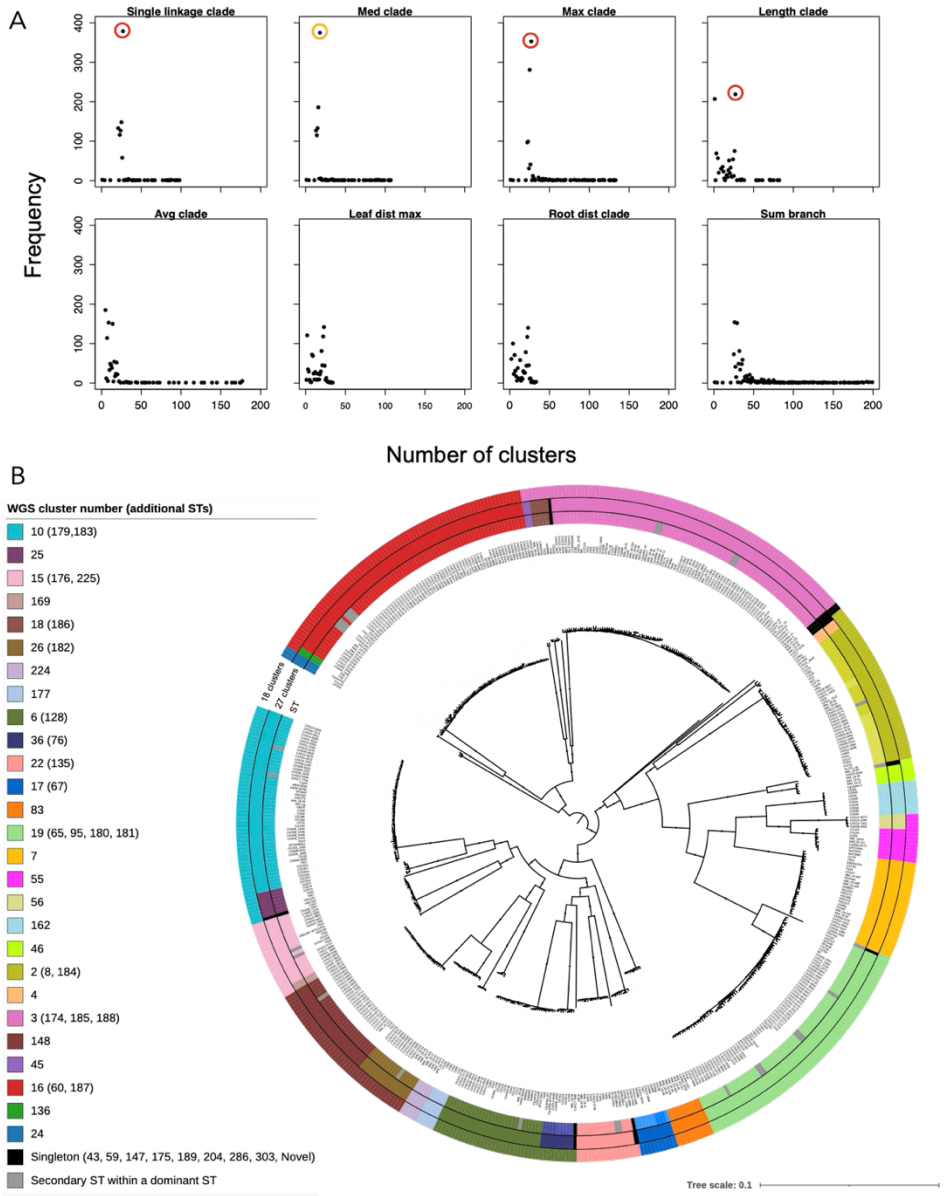


Figure 3.2: Clustering consistency and phylogenetic structure of 548 *N. glabratus* isolates. (A) Cluster prediction results using TreeCluster across eight different strategies, showing the number of clusters identified across 1000 distance thresholds. Three strategies (single linkage clade, max clade and length clade) identified 27 clusters across many thresholds (red circles). The med clade strategy also had a considerable threshold stability with 18 clusters (orange circle). (B) Maximum likelihood phylogeny of the 548 isolates. The inner ring shows sequence type (ST) assignments from the pubMLST database. The middle and outer rings display the 27 and 18 cluster designations respectively. In clusters containing multiple STs, the number listed in the legend is the predominant ST (which we propose using as the WGS cluster name). When multiple ST numbers are listed on the same row, they cluster together in the WGS phylogeny with the proposed WGS cluster number, i.e., the predominant ST group, listed first Black bubbles on branches represent bootstrap support ≥ 0.9.

The WGS isolates fell into 57 STs, and the sequence type distribution matched closely with the 27 cluster isolate designations from TreeCluster (Figure 3.2B). Fifteen of the 27 clusters contained isolates from a single ST group. Twelve (12) of the fourteen remaining clusters contained isolates from one dominant ST group alongside single or a small number of isolates from secondary STs; in all cases, isolates from the secondary STs genetically differed from the predominant ST MLST scheme in only one of the six genes, typically by a single allele (Table S3.2). Most of the MLST differences between isolates in dominant and secondary ST groups that clustered closely together occurred in *NMT1* (14/31) and *LEU2* loci (8/31). The other four genes were variable in fewer than five isolates (Table S3.2). Three clusters (ST2/ST8, ST17/ST67 and ST36/ST76) contained two dominant STs contributing an approximately equal number of isolates; each differed in only one gene (UGP1, TRP1 and LEU2, respectively). Nine isolates, all with unique STs, were singletons in the WGS phylogeny. The 18 cluster isolate designation consistently lumped smaller ST/cluster groups into adjacent larger isolate groups. From this point onward, the 27 clusters are each referred to by their dominant ST group. When there were two dominant groups, we compared the number of isolates on pubMLST and use the number with most isolates.

The 27 clusters are deeply branching from one another, explaining why MLST analysis does such a good job of recapitulating the WGS phylogeny. Genetic distance between the two most distant clusters (ST3 and ST56: 21 SNPs/kb) is an order of magnitude higher than that between the most closely related ones (cluster 18 and cluster 26 and cluster 55 and 56: 4 SNPs/kb). The genetic distance between closely related clusters are still two orders of magnitude larger than the genetic divergence within clusters (0.04 - 1.11 SNPs/kb for all clusters) (Table S3.3). Comparatively, the level of variation between closely-related *N.*

glabratus clusters is higher than the amount of genetic variation among distant clades in *C. albicans* (average of 3.7 SNPs/kb) (Hirakawa et al., 2015).

3.3.2 Admixture analyses

In unsupervised model-based clustering using ADMIXTURE, the cross-validation error was lowest at K = 27 (Figure 3.3A), consistent with the TreeCluster analysis. To assess the number of ancestral populations contributing to each isolate, we counted all ancestry components contributing 1% or more of the genome (i.e., $q \geq 0.01$). Isolates were classified as those with a single ancestry when a single component contributed at least 99% of the genome; 483 out of 548 isolates were single ancestry. We identified 65 admixed isolates (Figure 3.3B, Table S3.4) representing ~12% of all isolates. To minimize the impact of background noise, only ancestral proportions greater than 1% were counted. The majority (50) had contributions from two ancestral populations. Three isolates had ancestry from three populations, two from four, and one from five. The remaining nine isolates had contributions from seven or more populations. Across all admixed populations, fourteen isolates had a dominant ancestry component contributing more than 75% of their genome, while 32 isolates had the major component contributing between 50% and 75%. Thus, the vast majority of isolates derive their ancestry predominantly from a single population, while a small subset exhibits more complex patterns of admixture involving multiple ancestral sources.

All isolates within eighteen phylogenetically dispersed clusters shared only a single ancestor (Figure 3.3B). Three clusters (16, 19, 136) were made up of isolates from two different ancestries. Cluster 3 was made of three ancestral populations with isolates with

mixed ancestry. Cluster and 45 isolates had 4-5 mixed ancestries. Clusters 169 was particularly highly admixed, made up of over 10 ancestries. We also identified that clusters 55 and 56 are from the same ancestry, although the two differ in two allele profiles (*LEU2* and *URA3*).

Evidence of admixture was not equally detected among clusters. The majority of admixed isolates were from cluster 3 (32 isolates) and cluster 19 (14 isolates), with small numbers from cluster 16 (6 isolates), cluster 136 (2 isolates), cluster 169 (2 isolates), cluster 45 (2 isolates), and nine singletons had admixed ancestries. The regional variation in admixed isolates generally followed the sampling bias, with the majority originating from North America (31 isolates), followed by Europe (17), Oceania (5), and Asia (2). Admixed isolates were recovered from diverse clinical sources, again following sampling bias; most are from the circulatory system (33 isolates), with other isolation sites including unknown sources (13), abdominal samples (3), oral sites (3), urogenital tract (2), and the gastrointestinal tract (1 isolate). This distribution suggests that admixed lineages are not clinically restricted to specific body sites.

In addition to K = 27, we analyzed the ancestries of isolates at four other K-values (2,5,10, 16), selected based on the largest drops in the cross-validation scree plot (Figure S3.4). Of the 65 isolates from K =27 that were identified as admixed, 14 were of identified to be of mixed ancestry across all additional K-values, eight were admixed in one additional K-value, while 43 were not identified as admixed in the others. The consistently admixed isolates are from multiple clusters or are singletons (8/14), originated from diverse sources, and from diverse geographic regions.

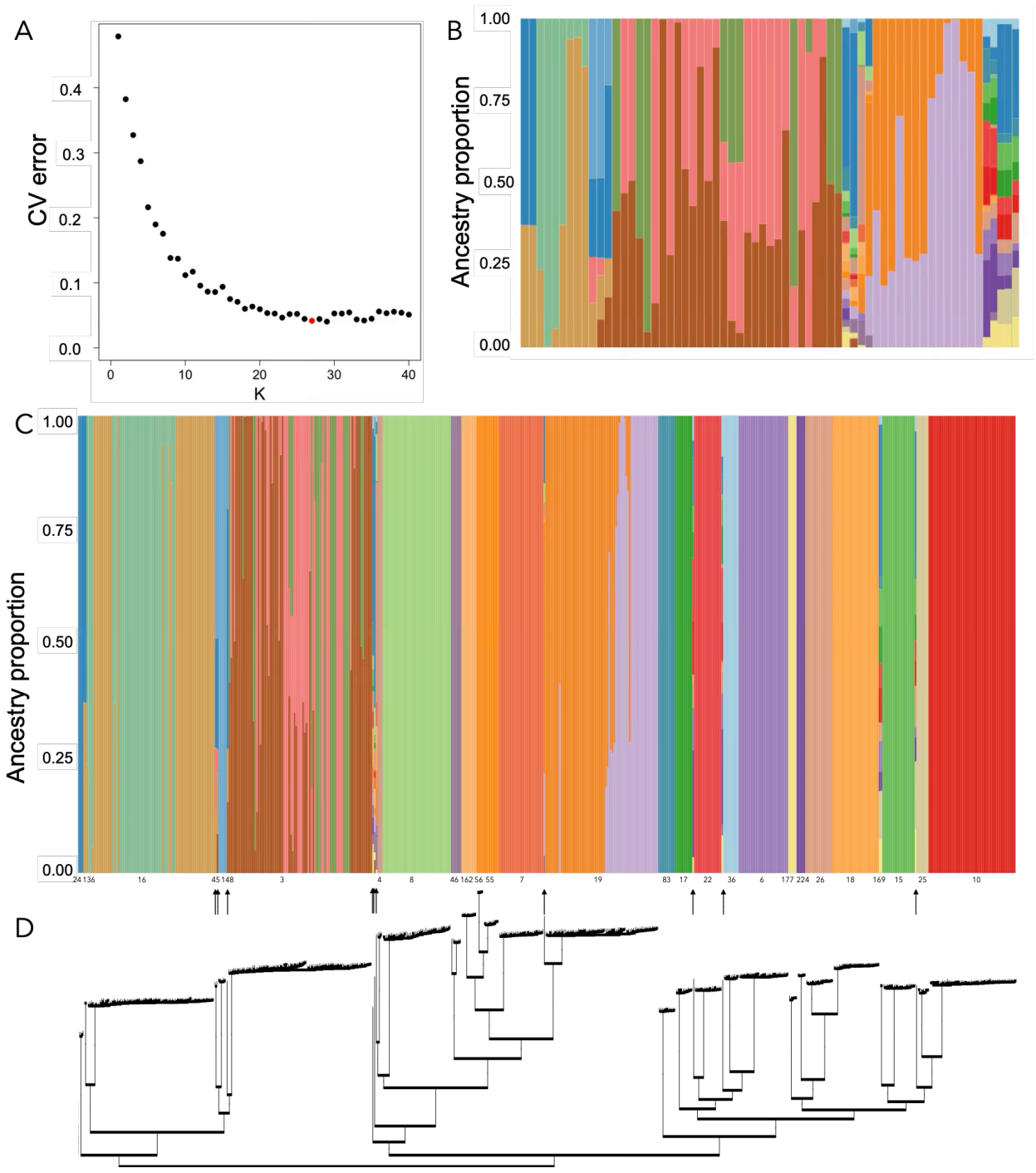


Figure 3.3: Population structure of *N. glabratus* inferred from genome-wide SNP data. A) Cross-validation (CV) error from unsupervised ADMIXTURE analysis of variant sites across the *N. glabratus* population, testing K-values from 1 to 40. The lowest CV error was observed at K = 27 and 29. B) ADMIXTURE plot for isolates that contain multiple ancestries. C) ADMIXTURE plot of all isolates at K = 27. Isolates are ordered according to the maximum likelihood phylogeny constructed with FastTree (D). Arrows indicate highly admixed isolates that appear as singletons in the phylogeny.

3.3.3 Geographical and isolation source structure

We found that the cluster 136 were exclusively identified in Asia (Table 3.1). In Europe, the overrepresented clusters were clusters 6, 22, 24, 148, and 177. In North America, clusters 16, 18, and 19 were overrepresented. In Oceania, clusters 83 and 26 showed significant overrepresentation. These findings suggest a strong geographic association for specific STs, indicating possible regional adaptation or localized transmission patterns.

Table 3.2: Distribution of *N. glabratus* clades across continents.

	ST3	ST4	ST6	ST7	ST8	ST10	ST15	ST16	ST17	ST18	ST19	ST22	ST24	ST25	ST26	ST36	ST45	ST46	ST55	ST56	ST83	ST136	ST148	ST162	ST169	ST177	ST224	S
Asia	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
Europe	29	2	17	8	15	20	6	0	5	2	12	12	3	5	0	6	0	1	6	0	1	0	5	9	0	5	1	2
North America	46	1	11	13	23	30	13	72	5	24	55	2	0	2	11	2	1	1	1	4	2	0	0	0	2	0	0	2
Oceania	8	0	1	5	2	1	0	3	0	1	0	2	0	0	5	1	1	4	2	0	7	0	0	0	0	0	4	3

3.3.4 Karyotypic variation

Aneuploidy was detected in 22 (4 %) isolates, with disomies observed across six different chromosomes (Figure 3.4A). No aneuploidies were observed for seven of the 13 chromosomes (B, G, H, I, J, K, M). The most frequently gained chromosome was chrE, present in 14 isolates. Additional recurrent disomies included chrA and chrD, each detected in 3 isolates, chrC, found in 2 isolates, and disomies of chrF and chrL, each observed in one isolate. While most isolates exhibited a gain in a single chromosome, three isolates showed multiple aneuploidies: one with disomies in chrC and chrE, one with disomies in chrA and chrD, and one with disomies in chrA, chrC, and chrE. To genomically assess whether aneuploid chromosomes are likely to be stable, we compared heterozygosity levels in diploid calling mode between aneuploid and non-aneuploid chromosomes. Similar to previous studies (Carreté et al., 2018), aneuploid chromosomes exhibited a similar number of predicted heterozygous SNPs compared to euploid haploid chromosomes, where heterozygous base

calls reflect a variant-calling error (Figure 3.5). This indicates that most aneuploid chromosomes in the *N. glabratus* WGS isolate set are recent, i.e., substantial *de novo* heterozygous variation has not accumulated.

Copy number variation (CNV) was also detected in several isolates, including five with aneuploidy (Figure 4A and 4B). The CNV ranged from 30kb to 810kb. One isolate had a 35kb segmental deletion on chrI. CNVs in chrM were found in three isolates with chrE aneuploidy and in one isolate with chrD disomy. Additional CNVs were observed in chrE, chrD, chrL, and chrK, with some involving large segmental amplifications. Four isolates exhibited CNVs across multiple chromosomes. Karyotypic variation was not thus restricted to a single clade (Figure S3.2).

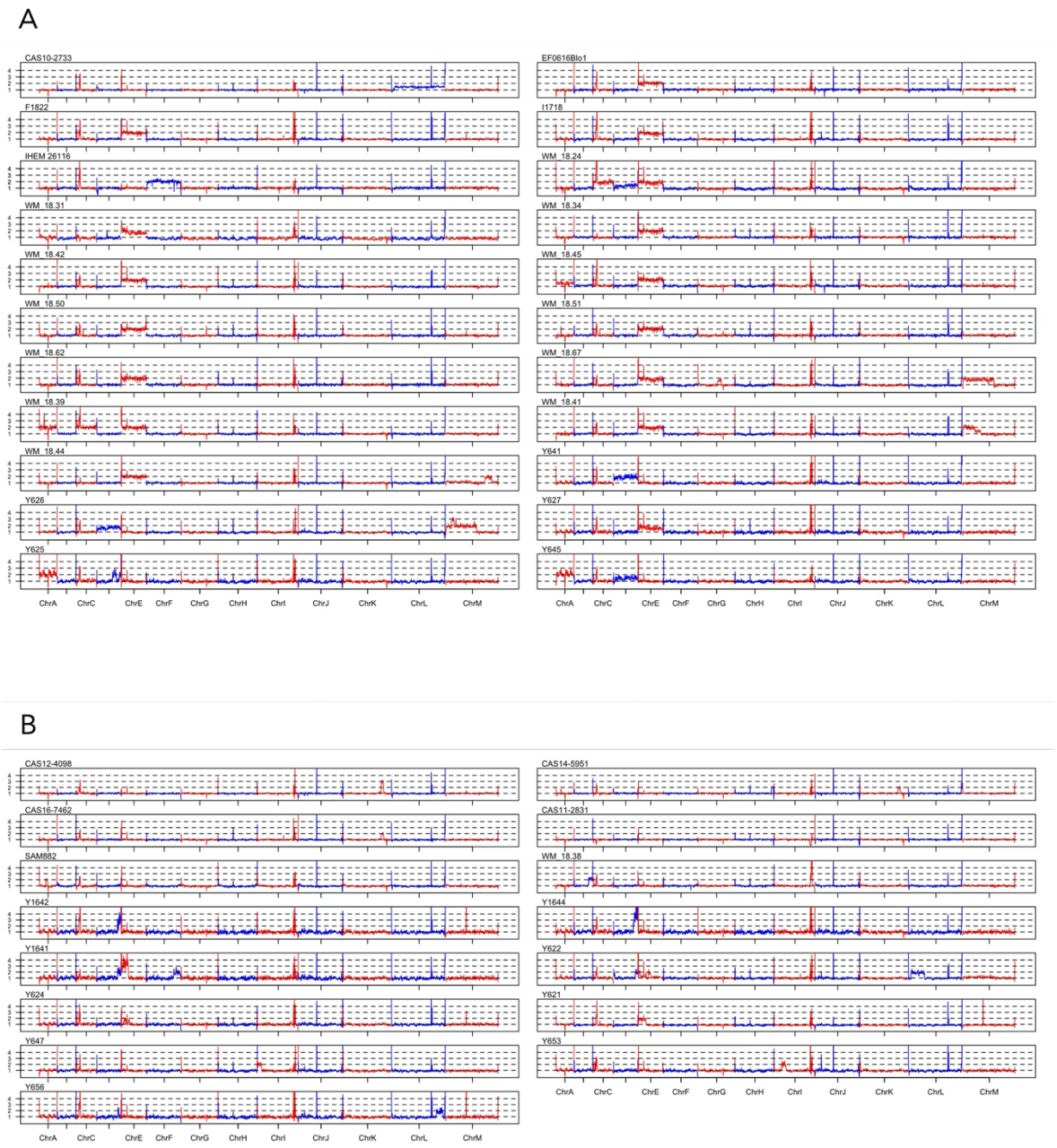


Figure 3.4: Whole-chromosome karyotypic variation among *N. glabratus* isolates.
Coverage was estimated using BEDTools genomecov, and average read depth was calculated over 5 kb sliding windows per chromosome using a custom R script. Each panel shows normalized coverage per chromosome for an individual isolate, highlighting variation in chromosome copy number. Chromosomes A–M are shown from left to right, and dashed horizontal lines indicate ploidy levels (e.g., 1×, 2×, 3×). (A) Aneuploid isolates, showing whole-chromosome copy number variation. (B) Isolates with segmental copy number variation (CNV). Note that some aneuploid isolates also exhibit CNV.

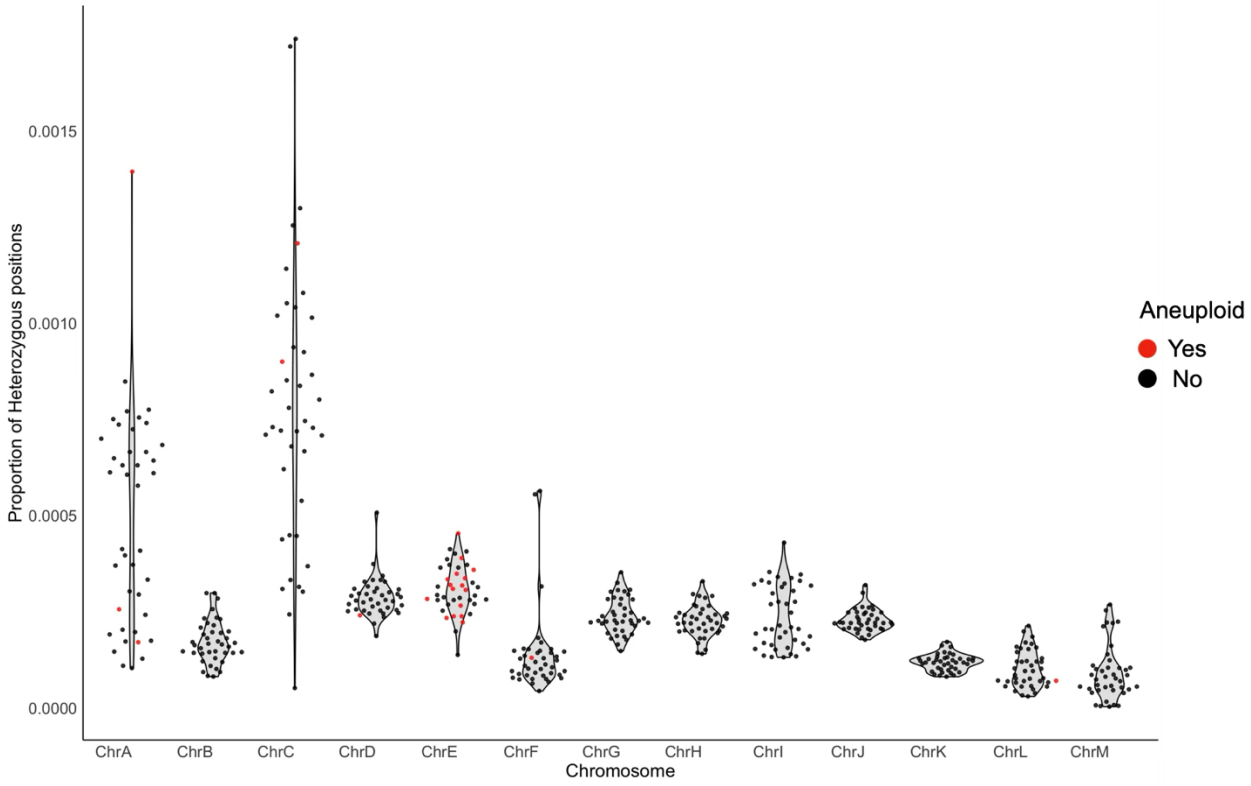


Figure 2: Distribution of homozygous and heterozygous SNPs in aneuploid isolates called in diploid mode. Levels of heterozygosity per chromosome are similar for both aneuploid and non-aneuploid isolates.

3.4 Discussion

Our analysis of 548 isolates revealed that WGS-defined clusters correspond closely with MLST-defined topology. Fourteen WGS clusters consist exclusively of a single ST, while mixed clusters typically include a dominant ST and one or a few closely related STs with few variants differentiating them. Given these close genetic relationships, we propose that clusters containing multiple STs be labelled according to the dominant ST. This approach aligns with pragmatic practices in microbial genomics, such as in *Listeria monocytogenes*, where clonal complexes are defined as groups of STs differing by one allele, and the "complex" takes the name of the largest or first-described ST (Chenal-Francisque et al., 2011; Haase et al., 2014; Ragon et al., 2008). A similar designation system is also used in *Staphylococcus aureus* (Feil et al., 2003), where clonal complexes are defined as groups of STs that match a central genotype at four or more loci, unless they are more closely related to a different central genotype (https://pubmlst.org/organisms/staphylococcus-aureus/clonal-complexes). Such naming promotes clarity and facilitates communication across studies.

This approach is different from what has previously been done in *N. glabratus* WGS studies. Dodgson et al. (2003) and Carreté et al. (2018) both advocated for the use of Roman-numeral cluster naming, while Helmstetter et al. (2022) supported ST-based nomenclature tied strictly to MLST profiles. The naming strategy proposed here strikes a balanced middle ground. By retaining cluster names based on the primary ST's label, we reconcile precision with practicality. Within clusters containing multiple STs, we found the NMT1 locus to be the most variable. This aligns with previous observations highlighting NMT1 as the most diverse among the six loci used in the MLST scheme (Dodgson et al., 2003; Gabaldón et al., 2020).

We identified 65 isolates with evidence of admixture, each showing contributions from various ancestral populations. Most admixed isolates were within cluster 3, a cluster previously associated with admixture (Carreté et al., 2018), while 13 admixed isolates were found in cluster 19, which had not been previously identified as mixed-ancestry. In analyzing 181 isolates, Helmstetter et al. (2022) identified five admixed isolates, three of which were confirmed in this study (CG57, WM_18.66, WM_05.155). Of the two isolates remaining, one (CG1), was not detected as admixed in our analyses. The other isolate, M17, was not available for download at the time of the analyses. The number of ancestral populations detected in admixed isolates ranged from 2 to 18, supporting the view that recombination in *N. glabratus* is not only an ancient feature of its evolutionary history but also a process that continues to shape its population structure. Admixed isolates were geographically widespread, which may serve to maintain or enhance genetic diversity within *N. glabratus* (Steensels et al., 2021). A peculiar observation in this study as well as other was the observation that the admixture K used as optimal was closely similar to the number of clusters observed in the phylogeny (Carreté et al. 2018; Helmstetter et al. 2022). This is most likely because of the violation of loci independence, which is contrary to the assumptions underlying admixture analyses. Through linkage disequilibrium analysis Xu et al. (2025) estimated from the six MLST alleles, a positive standardized index of association value of 0.2492 and a variance of pairwise differences (2.0530) > linkage equilibrium (0.9989) indicating the population is clonal population structure and LD exists in N. glabratus. We are thus very cautious in over-interpreting the results. The rarity and small size of these admixed groups indicate minimal introgression and restricted gene flow among major clusters. This limited admixture, together with the strong genetic differentiation and deeply branching phylogenetic

structure, further supports the hypothesis that the current N. glabratus isolate collection could represents a species complex.

Our findings highlight the influence of geographic sampling bias on the observed distribution of *N. glabratus* STs. While we identified certain STs as overrepresented in specific regions, discrepancies with previous studies suggest that such patterns may be shaped partly by true geographic structuring and party by sampling bias. Our identification of ST16 as overrepresented in North America aligns with others (Kuthan, 2025), as does the broad distribution of ST3 across multiple continents. However, while Dodgson et al. (2003) reported an enrichment of ST19 in Europe, our study and that of Biswas et al. (2018) found it to be more prevalent in North America. Similarly, the absence of Asian ST7 isolates in our study (previously described as enriched in Japan, (Dodgson et al., 2003)) likely reflects our lack of sequencing data from that region, rather than the true absence of ST7. Combined, these observations emphasize the need for more geographically balanced sampling and sequencing efforts. Current datasets, including our own and the pubMLST dataset for *N. glabratus*, remain skewed toward North America and Europe, limiting our ability to draw firm conclusions about global primary ST distribution. A concerted effort to include underrepresented regions will be essential to resolve conflicting findings and to better understand the population structure of *N. glabratus* populations.

Aneuploidy was detected in 22 (4 %) isolates, which is consistent with others (Zheng et al., 2022). ChrE, carries the ERG11 gene (the target of azole drugs) and its amplification has been shown to correlate with resistance phenotypes (Marichal et al., 1997). This is similar to *C. albicans* where amplification of chromosome 5 (which harbours the *ERG11* and *TAC1* loci) confers a reversible azole resistance (Selmecki et al., 2009; Vande Zande et al.,

2023). Our data in *N. glabratus* also show that aneuploid chromosomes lack elevated heterozygosity, which supports their classification as recent *de novo* events (Carreté et al., 2018). This pattern is mirrored in other studies where aneuploidy emerges under stress and reverts during relaxed conditions. Furthermore, the presence of CNVs (both segmental deletions and amplifications) expanding across several chromosomes and co-occurring with aneuploidies denotes another layer of genomic plasticity. In fungal systems, CNVs are commonly implicated in stress response and adaptive evolution (Jay et al., 2025); for example, *C. albicans* develops isochromosomes or segmental amplifications under antifungal exposure, and CNVs are known to arise via multiple mechanisms under cellular stress (Todd and Selmecki, 2020; Vande Zande et al., 2023).

3.6 Conclusion

Taken together, our findings demonstrate that while MLST remains broadly concordant with WGS-based phylogenetic structure, whole-genome data can detect and resolve fine-scale sequence type nuances within *N. glabratus* phylogenies. The proposed ST-based cluster naming provides a pragmatic alternative to previous inconsistent nomenclatures and is supported by practices in other microbial systems. Our identification of widespread admixture, deep phylogenetic divergence, and structural genome variation (including aneuploidies and CNVs) indicates that *N. glabratus* harbors extensive genomic diversity shaped by both recombination and karyotypic plasticity.

Chapter 4: Migration and standing variation in vaginal and rectal yeast populations in recurrent vulvovaginal candidiasis

This chapter has been published as a peer-reviewed journal article:

Adamu Bukari, A. R., Kukurudz, R. J., de Graaf A., Habon D., Manyaz B., Syvolos Y., Sumanarathnea A., Poliquin V., and Gerstein, A. C. (2025). Migration and standing variation in vaginal and rectal yeast populations in recurrent vulvovaginal candidiasis. mSystems, doi: https://doi.org/10.1128/msystems.00157-25

American Society for Microbiology Journal permits authors to use published content for a thesis.

https://journals.asm.org/author-self-archiving-permissions

4.1 Abstract

Vulvovaginal candidiasis is a very common human fungal infection. Most are successfully treated with antifungal drugs, yet ~8% lead to recurrent vulvovaginal candidiasis ("RVVC"). Vaginal and rectal populations have been previously found to be closely related in RVVC. Yet the genomic methods used precluded the assessment of fine-scale relationships and the measurement of within-population variation, a fundamental property with evolutionary implications. To address this gap, we isolated 12 vaginal and 12 rectal yeast isolates from each of four individuals with a history of RVVC. Three individuals had *Candida albicans* infections, while the fourth had *Nakaseomyces glabratus*. All isolates were whole-genome sequenced and phenotyped. Isolates were placed into species-level phylogenies composed of isolates from many different countries and contexts, including an updated *N. glabratus* tree with over 500 isolates. Genotypic and phenotypic analyses were consistent with migration between sites. There was little phenotypic diversity in drug response and no consistent difference between isolates from different sites for invasive growth. Although there are few comparables, *C. albicans* nucleotide diversity was similar to most commensal oral and rectal populations, while *N. glabratus* was similar to some bloodstream infections (though higher than others). Single-nucleotide changes drove intra-population genetic differences; only a single loss-of-heterozygosity tract varied among isolates from within one participant. This study provides baseline measurements and describes techniques to quantify within-population diversity in fungal microbes. We highlight a need for comparable studies that use the same sampling effort and analysis methods to understand the interplay between evolutionary factors in shaping fungal microbial communities.

4.2 Introduction

Vulvovaginal candidiasis (VVC, colloquially "yeast infection") is common, affecting approximately 75% of people defined female at birth at least once in their lives (Benedict et al., 2022; Rathod and Buffler, 2014; Yano et al., 2019). The disease burden of VVC results in global annual treatment costs of ~ 1.8 billion U.S. dollars (Denning et al., 2018; Foxman et al., 2000), with a loss in productivity in high-income countries of ~ 14 billion U.S. dollars (Denning et al., 2018). Treatment involves topical or oral antifungal medication, which is effective at symptom abatement in most cases. However, ~ 8% of individuals with VVC experience recurrence (RVVC), defined as three or more symptomatic episodes a year (Denning et al., 2018; Foxman et al., 2000). Co-morbidities that involve the vaginal microbiome, such as recurrent bacterial vaginosis (Benyas and Sobel, 2022) or frequently taking antibiotics to treat conditions such as cystic fibrosis (Kazmerski et al., 2018) are known to predispose individuals to RVVC. Similarly, treatment for conditions that lead to alteration of the vaginal mucosa, such as hormone replacement therapy in postmenopausal women (Fischer and Bradford, 2011) has also been associated with increased prevalence, as has a small number of human genetic variants such as mannose-binding lectin deficiency (Babula et al., 2003) and TLR2 Pro631His polymorphism (Rosentul et al., 2014). Yet approximately half of all people with RVVC have no identifiable risk factors (Sobel, 2003), signifying the need for additional studies on the biological basis of this chronic condition.

Candida albicans is responsible for 50-90% of VVC and RVVC cases (collectively, R/VVC) (Guzel et al., 2011; Shi et al., 2015; Song et al., 2022; van Schalkwyk et al., 2015; Zhang et al., 2014). *Nakaseomyces glabratus* (formerly *Candida glabrata*) is the second most prevalent cause, globally attributed to ~ 8% of cases (Kennedy and Sobel, 2010; Parazzini et

al., 2000; Richter et al., 2005). Here, we collectively refer to these species using the colloquial term "yeast," which reflects a shared morphology while acknowledging our current understanding of their phylogenetic relationships and the recent official renaming of *N. glabratus* (Borman and Johnson, 2021; Kidd et al., 2023). To be consistent with clinical practice, we continue to use the R/VVC abbreviations while noting that "candidiasis" does not reflect the updated genera names. Treatment recommendation for R/VVC differs by species, as many isolates from *N. glabratus* (and other non-*albicans* pathogenic yeast species) have intrinsic resistance to the azole antifungal fluconazole that is commonly used to treat RVVC (Pfaller et al., 2015).

Understanding the etiology of R/VVC is complicated in part since yeasts are a common commensal member of the vaginal microbiota without causing symptomatic VVC (Drell et al., 2013), and *C. albicans* studies repeatedly find no strict phylogenetic differentiation between commensal and pathogenic strains (Ropars et al., 2018). Relapse in RVVC, i.e., return of symptoms, could theoretically be due to either incomplete eradication of the vaginal yeast population after taking antifungal drugs or complete vaginal eradication followed by re-colonization (Tasić et al., 2002). Decades of studies have sought to understand the etiology of RVVC, as the answer has potential implications for improving treatment and reducing or eliminating symptom recurrence. The GI tract has been suggested as a possible endogenous source population, yet a study in 1979 that treated RVVC patients with oral nystatin to reduce the resident GI population found that it did not decrease the time to recurrence (Milne and Warnock, 1979). Furthermore, studies examining yeast colonization of the GI tract through feces or rectal swabs during symptomatic recurrence find that not all participants are culture-positive (El-Din et al., 2001; Fong, 1994; Mårdh et

al., 2003; Milne and Warnock, 1979; O'Connor and Sobel, 1986; Sobel, 1986; Spinillo et al., 1994). However, this does not necessarily preclude that a small GI population is present in all individuals (below the culture detection limit in some), which could act as an endogenous reintroduction source under the right host conditions. Examining the diversity of strains at different body sites and among recurrent infections can potentially differentiate between relapse scenarios. If vaginal isolates are closely related to GI tract/rectal isolates but less diverse, that would be consistent with reintroduction. If vaginal isolates sampled at different time points consistently have the same genotype that is not present in the GI tract/rectal isolates, this would be consistent with incomplete eradication. Looking at the level of diversity and relationships among isolates acquired within a single time point ("standing variation") at different body sites can also help us understand the adaptive potential and migration dynamics of yeast populations within the body.

Multilocus sequence typing (MLST) has been commonly used for phylogenetic analyses in the context of R/VVC (Song et al., 2022; Tian et al., 2021; Zhu et al., 2022). The overarching results are generally consistent with the maintenance of genotypes between symptomatic recurrences, with a few examples of novel genotypes arising at one time point compared to another. However, typically only one or a small number of isolates are examined at a given time. Thus, standing genetic variation could have been undetected. Only a single study sequenced two vaginal isolates from different time points using short-read whole genome sequencing (WGS), and no previous studies have employed WGS to compare multiple isolates from the same time point. WGS removes the need to rely on a single or a small number of markers, which could over or underinflate the actual level of diversity. For example, Sitterlé et al. (2019) showed that while MLST revealed occasional differences

among *C. albicans* oral isolates collected from three healthy individuals, whole genome sequencing revealed that the three examined isolates were actually closely related in each case.

Standing genetic variation of yeast populations has only been quantified in a handful of contexts. Two studies in *C. albicans* that whole genome sequenced 3-6 oral and rectal isolates from six healthy individuals found that although isolates were closely related in most cases, they differed by numerous single-nucleotide polymorphisms, primarily resulting from short-range loss-of-heterozygosity tracts (Anderson et al., 2023; Sitterlé et al., 2019). Two individuals were, however, simultaneously colonized with oral isolates from different clusters (Anderson et al., 2023). A single *N. glabratus* study sequenced up to ten isolates from nine patients with bloodstream candidemia (Badrane et al., 2023); a pairwise SNP analysis was also consistent with closely related isolates. Variation in RVVC has yet to be quantified through WGS; hence, whether it is similar to commensal populations is unknown. It has also not been determined how many isolates must be sequenced to capture population diversity accurately.

Here, we build on previous work by conducting WGS paired with high-throughput phenotyping to quantify vaginal and rectal standing variation in participants with a history of RVVC when high vaginal and rectal yeast population sizes were observed. We compared our results to the few comparable studies that conducted WGS of contemporaneous yeast isolates from other contexts and statistically evaluated the minimum number of isolates required to accurately measure genetic variation. We obtained isolates from the same time point from four individuals with a history of RVVC. Three participants were infected with *C. albicans*, and one with *N. glabratus*. From all individuals, we found a complete phylogenetic

overlap of vaginal and rectal isolates and no consistent difference in phenotypes, consistent with high levels of migration. We found no evidence that diversity in populations from the two sites was different; levels of standing genetic variation were generally similar to what has been observed in other contexts. This suggests that despite frequent population bottlenecks caused by drug treatment, vaginal yeast diversity is maintained or rapidly restored.

4.3 Materials and methods

4.3.1 Clinical isolates

Seventeen consenting female participants who attended a clinic for individuals with a history of RVVC in Winnipeg, Canada, were sampled at the clinic for possible inclusion in this study. Vaginal and rectal swabs were acquired from all participants; swabs were kept at -4 °C or on ice during transport and processed within 5 h of acquisition. Swabs were agitated for ~30 s in 1 mL PBS, a dilution series (1, 1:10, 1:100) was conducted, and 100 µL from each dilution was spread onto SDA and chromogenic *Candida* agar plates (CHROMagar™ by Dalynn Biologicals). Plates were incubated for 48 h at 30 °C. To meet our goals of comparing variation between vaginal and rectal populations and to infer the number of isolates required to accurately measure genotypic diversity, we focused our efforts on the four participants who had high vaginal and rectal populations (> 1×10^3 CFU/mL of swab elute). Based on the colony color on chromogenic agar, one was *N. glabratus* (YST6) and three were *C. albicans* (YST7, TVY4, TVY10). Twelve vaginal and 12 rectal colonies with clear margins were haphazardly isolated from each participant, suspended in 1 mL of 20% glycerol, and kept at -70 °C. We collectively refer to these 96 isolates collected from the four of seventeen participants who met our inclusion criteria for this study as the "THRIVE-yeast" isolates. This study has been approved by the University of Manitoba Biomedical Research Ethics Board (HS24769 (B2021:026)) and Shared Health (SH2021:038).

4.3.2 DNA extraction and sequencing

Genomic DNA was extracted from all THRIVE-yeast isolates following a standard phenol-chloroform protocol as previously described (Kukurudz et al., 2022). DNA quality and concentration were assessed spectrophotometrically (NanoDrop 2000, Thermo Scientific™) and fluorometrically (Qubit® 2.0 Fluorometer). Genomic DNA was sent to either Microbial Genome Sequencing Center ("MIGS," Pittsburgh, USA; YST6 and YST7) or SeqCoast Genomics (New Hampshire Ave., USA; TVY4 and TVY10) for sequencing. At MIGS, sample libraries were prepared using the Illumina DNA Prep kit and IDT 10bp UDI indices and sequenced on NextSeq 2000 using a 300-cycle flow cell kit, producing 2×151bp reads. The bcl-convert v3.9.3 software was used to assess read quality, demultiplex and trim adapter sequences. At SeqCoast Genomics, samples were prepared using an Illumina DNA Prep tagmentation kit and unique dual indexes. Sequencing was performed on the Illumina NextSeq2000 platform using a 300-cycle flow cell kit to produce 2×150bp paired reads. DRAGEN v3.10.11 was used to assess read quality, demultiplex and trim adapter sequences. Three of the 96 isolates (TVY10R13, TVY4R4 and YST7R13) had extremely low coverage (<20×) and were excluded from genomic analysis. The average coverage from the remaining 93 isolates was at least 50×. The fastq files from all THRIVE-yeast have been deposited at the National Center for Biotechnology Information (NCBI) Sequence Read Archive under BioProject ID PRJNA991137.

In addition to the 93 genomes we sequenced, we downloaded an additional 182 *C. albicans* FASTQ files from the NCBI Sequence Read Archive database (Sayers et al., 2022) from BioProject Accession PRJNA432884 (Ropars et al., 2018) and 526 *N. glabratus* FASTQ files from 19 different projects on the SRA database (assessed on February 05, 2022),

including 99 *N. glabratus* FASTQ files from SRA PRJNA361477 (Carreté et al., 2018) and PRJNA669061 (Helmstetter et al., 2022) (see Table S4.1 and Table S4.2 at https://github.com/microstatslab/RVVC/supporting_tables).

4.3.3 Variant calling

The sequence reads were trimmed with Trimmomatic (v0.39) (Bolger et al., 2014) with standard parameters (Todd et al., 2019). Quality was assessed with FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and MultiQC (Ewels et al., 2016). *C. albicans* trimmed paired-end reads were mapped using bwa-mem (Li, 2013) to the SC5314 haplotype A reference genome (A22-s07-m01-r160) downloaded from the Candida Genome Database (Skrzypek et al., 2017). The resulting SAM file was coordinate-sorted and converted to a Binary Alignment Map (BAM) file using samtools v1.9 (Li et al., 2009). *N. glabratus* isolates were mapped to the CBS 138 reference genome (GCA000002545v2) downloaded from the Ensembl Genome Database (Yates et al., 2022). Alignment quality was assessed with CollectAlignmentSummaryMetrics from Picard v2.26.3 (http://broadinstitute.github.io/picard) and consolidated across all samples with MultiQC (Ewels et al., 2016). All files had a >95% mapping quality. BAM files were further processed with Picard by adding a read group annotation so that samples with the same BioProject ID had the same read group, removing duplicate PCR amplicons and fixing mate pairs. Base quality scores for the *C. albicans* aligned reads were recalibrated with known single-nucleotide polymorphisms obtained from the Candida Genome Database website (http://www.candidagenome.org/download/gff/C_albicans_SC5314/Assembly22/A22_Jones_PMID_15123810_Polymorphisms.vcf; downloaded on July 29, 2020) (Jones et al., 2004)

using the BaseRecalibrator and ApplyBQSR from the Genome Analysis Toolkit 4.2.4.0. The average coverage for each isolate was estimated using samtools v1.9 (Li, 2013).

The GATK Best Practices were adapted for variant calling. In sequence, HaplotypeCaller, CombineGVCFs, GenotypeVCFs, VariantFiltration, and SelectVariants (DePristo et al., 2011; Poplin et al., 2018; Van der Auwera et al., 2013) were used to identify single nucleotide variants (SNPs) among all sequenced isolates in diploid and haploid mode for *C. albicans* and *N. glabratus*, respectively. The resulting SNP table was hard filtered using the suggested parameters and to match (Ropars et al., 2018) (QualByDepth < 2.0, FisherStrand > 60.0, root mean square mapping quality < 30.0, MappingQualityRankSumTest < −12.5, ReadPosRankSumTest < −8.0). We excluded variants that were called in known repetitive regions of the genome, as these are likely to reflect sequencing misalignments rather than true variants, i.e., the subtelomeric regions (15kb from the start and end of each chromosome), the centromeres, and the major repeat sequence regions (Table S4.3, start and stop positions from candidagenome.org).

4.3.4 Phylogeny construction

Phylogenetic trees were constructed for *C. albicans* and *N. glabratus*. For each species, the multi-sample VCF file consisting of genomic SNPs was converted to a FASTA alignment using vcf2phylip.py v2.8 (Ortiz, 2019). For heterozygous SNPs in *C. albicans,* the consensus sequence was preferentially made based on the reference (haplotype A) base. Ambiguous bases were written following IUPAC nucleotide ambiguity codes in the matrix. The FASTA alignment was parsed in FastTree (2.1.11) (Price et al., 2010) in the double precision mode to construct an approximate maximum-likelihood phylogenetic tree using the general time

reversible model and the -gamma option (GTR+G) to rescale the branch lengths. The phylogeny was visualized and annotated with the Interactive Tree Of Life (iTOL, v5) (Letunic and Bork, 2021). Isolates from *C. albicans* clade 13, i.e., *C. africana* (Mixão et al., 2021; Mixão and Gabaldón, 2020; Romeo et al., 2013), were used to root the phylogeny. The *N. glabratus* phylogeny was rooted at the midpoint. Following standard practice for *N. glabratus*, we used MLST analysis (Carreté et al., 2018; Helmstetter et al., 2022) to identify and name the YST6 ST group. An *in silico* MLST analysis from the fastq data was done using stringMLST (Gupta et al., 2017) using the predefined MLST *N. glabratus* allele library (*FKS*, *LEU2*, *NMT1*, *TRP1*, *UGP1*, *URA3*) from the PubMLST database (Jolley et al., 2018).

To phylogenetically compare the vaginal and rectal isolates within each participant, we constructed a maximum likelihood phylogeny with RAxML v8.2.12 following standard practice using the GTR+G model with 20 ML inferences on the alignment, inferring bootstrap replicate trees, applying MRE-based bootstrapping test, and drawing support values using Transfer Bootstrap Expectation on the best-scoring tree (Stamatakis, 2014).

4.3.5 Diversity analysis of isolates from each participant

Real time genomics (RTG)tools with the vcfsplit option (Cleary et al., 2015) was used to extract a VCF files for each isolate from the multi-sample VCF file. The VCF files were converted to BED files using "bcftools query -f,". A custom R script was used to do a pairwise comparison of the isolates from a site to determine differences in SNP positions ("05a_SNP_difference.R" and "05b_SNP_difference_plotStats.R"). An ANOVA test was performed to compare pairwise SNP differences in rectal and vaginal isolates from each participant.

A principal component analysis on windows of genomic regions that differed among isolates from the same participant was conducted. Using the multi-sample VCF file of SNPs for each participant, the "templated script" from the R package lostruct [local PCA/population structure, v.0.0.0.9, (Li and Ralph, 2019)] was with run with parameters -t: bp, -s: 5000, -npc: 2 and -m: 2. To check for possible segregation of vaginal and rectal isolates, plots were generated based on the most extreme heterogeneous genomic windows across the entire genome generated as part of the lostruct pipeline.

4.3.6 Relationship between average nucleotide diversity (π) and number of samples

The potential influence of the sample size was assessed using the average pairwise diversity differences between all possible isolate pair estimates from Pixy (v1.2.6.beta1), (Korunes and Samuk, 2021). Briefly, variants were called using GenotypeGVCFs with --all-sites option activated. Vcftools (v0.1.16) (Danecek et al., 2011) was used to filter the variants (with --max-meanDP 500, --min-meanDP 20, --max-missing 0.8). Indels and mitochondrial DNA were excluded (--remove-indels, --not-chr). For each sample size, we randomly selected *n* vaginal isolates 50 times without replacement, and calculated π for batch sample using Pixy (v1.2.6.beta1).

4.3.7 *In silico* mating-type locus detection

To determine mating type-like locus in the *C. albicans* isolates (YST7, TVY4 and TVY10), the reads were aligned to both haplotypes and consensus sequences of the MAT locus on chromosome 5 (*MATa1* and *MATa2* for hapA and *MATα1* and *MATα2*) were extracted. A BLAST search was then conducted to confirm the locus. Similarly, the mating

type-like locus of *N. glabratus* (YST6) isolates was determined by determining the consensus sequence for MTL1 (*MATα1* and *MATα2*) and MTL3 on chromosome B, and MTL2 on chromosome E and confirming the MTL by a BLAST search.

4.3.8 Loss of heterozygosity and copy number analyses

Loss of heterozygosity (LOH) and copy number variant (CNV) analyses were conducted using the web-based yeast analysis pipeline (Y_{MAP}) (Abbey et al., 2014). *N. glabratus* isolates were compared to the CBS138 reference genome (CGD: s05-m01-r09). *C. albicans* isolate were analyzed against the SC5314 A22-s02-m09-r10 reference genome. Correction was enabled for GC-content bias and chromosome-end bias. The genomic elements within observed CNV regions were identified using the "gene/sequence resources" section of the Candida Genome Database (CGD, https://candidagenome.org).

4.3.9 Growth Rate Assay

Two separate growth rate assays were conducted to measure growth in RPMI (10.4 % w/v RPMI powder, 1.5 % w/v dextrose, 1.73 % w/v 3-(N-morpholino) propane sulfonic acid (MOPS), adjusted to pH 7 with NaOH tablets) and vaginal simulative medium ("VSM", following (Ropars et al., 2018): 1.16 % v/v 5 mM NaCl, 3.6 % v/v 0.5 M KOH, 0.0128 % v/v 99% glycerol, 20 % v/v 0.01 M $Ca(OH)_2$, 1.34 % v/v 0.5 M Urea, 6.6 % v/v 0.5 M glucose, 0.67% w/v solid Yeast Nitrogen Base, 0.85 % v/v 2 M acetic acid, 0.192 % v/v lactic acid, adjusted to pH 4.2 with NaOH tablets). For each, 5 µL of frozen glycerol stock from all THRIVE-yeast isolates was inoculated in duplicate into 500 µL of RPMI or VSM and incubated for 48 h at 37 °C with agitation at 250 rpm. Cultures were standardized to an optical density

(OD) of 0.01 A600 in RPMI or VSM, and 200 µL was transferred into a 96-well round bottom plate and sealed with a Breath-Easy sealing membrane (Electron Microscopy Sciences, PA, United States). OD_{600} readings were taken by the Epoch plate reader (Biotek) every 15 minutes, with orbital shaking at 37 °C for 48 h. From each well, the maximal growth rate was calculated as the spline with the highest slope using a custom R script written by Dr. Richard Fitzjohn (https://github.com/acgerstein/THRIVE-variation/scripts_real/growthrate_generic.R). The average growth rate between two technical replicates for each isolate in each growth medium was used for visualization and statistical analysis. Statistical outliers were determined through Rosner's test of outliers available through the rosnerTest function in the EnvStats R package (Millard, 2013). For each population, we started with a k value of one (i.e., testing for a single outlier). If that was significant, we increased k by one until no additional outliers were identified (i.e outliers were not excluded).

4.3.10 Drug resistance and tolerance

Disk diffusion assays were performed to quantify variation among isolates in resistance and tolerance. A pilot experiment was done on 24 isolates from YST7 in five different drugs (FLC: fluconazole, CLT: clotrimazole, MCZ: miconazole, NYT: nystatin, BA: boric acid). Subsequently, disk diffusion assays were conducted on all isolates to fluconazole and boric acid at pH 4.2. We chose to focus our efforts on fluconazole and boric acid as these are drugs in different classes that are both treatment options for induction and maintenance therapy of recurrent VVC and boric acid is used in treatment of non-albicans VVC (van Schalkwyk et al., 2015). Except the pH adjustment, the protocol outlined in the NCCLS M44

guidelines for antifungal disk diffusion susceptibility testing for fluconazole was followed (Rex and Clinical, 2009) and adapted for boric acid as previously described (Salama and Gerstein, 2022). The entire experiment was conducted twice for each isolate × drug, with two technical replicates for each of the two biological replicates. Previous work demonstrated consistent drug resistance values at 24 h and 48 h, with drug tolerance apparent at 48 h. Photographs were taken at 48 h images and processed in ImageJ as previously described (Salama and Gerstein, 2022) then run through the diskImageR package (Gerstein et al., 2016) for drug resistance (RAD_{20}) and tolerance (FOG_{20}) quantification. Briefly, diskImageR calculates resistance as RAD_{20}, as the radius of the zone of inhibition where growth is reduced by 20% relative to growth on the margins of the plate where there is no drug, and tolerance as FoG_{20}, the fraction of realized growth between RAD_{20} and the disk.

A Welch two-sample t-test that did not assume equal variance was used for each participant × drug combination to compare vaginal and rectal isolates. Statistical outliers were determined through Rosner's test of outliers for growth rates. All statistical analysis was done at a type I error rate of 0.05.

4.3.11 Invasive growth assay

To examine invasive growth, we revised methods from (Zupan and Raspor, 2008). Freezer stock from THRIVE-yeast isolates were streaked onto 20 mL yeast peptone dextrose (YPD) with additives plates (2% w/v peptone, 2% w/v yeast extract, 1.8% w/v agar, 1% w/v glucose, 0.00016% w/v adenine sulphate, 0.00008% w/v uridine, 0.1% v/v of chloramphenicol and ampicillin) and grown for 72 h at room temperature. A single colony

was then randomly chosen from each isolate and inoculated into 200 µL YPD broth. If no single colonies were available, a similar amount of culture from the colony lawn was used. Cultures were standardized to OD_{600} 0.01 in 1 mL of liquid YPD media, then 2 µL of standardized culture was spotted onto the surface of a 20 mL solid YPD plate in a hexagonal pattern for a total of 7 spots per plate (i.e., spotted at each vertex and in the center). Plates were incubated for 96 h at 37 °C. The surface growth was washed off using distilled water from a squeeze bottle and a photograph was taken in a dark room on a lightbox. Two biological replicates were performed for each isolate.

The qualitative amount of invasive growth for each isolate was determined by visual examination of the post-wash photographs. To develop a five-point scale, two different people independently went through the post-wash pictures from YST6 and YST7 and selected two to six representative pictures that fit into five levels of invasive growth (scored as 1-5). The independent selections were then compared, and one image from each person was chosen as the most representative for each level of the scale. Using these as a reference, each isolate was then categorized into the five levels of the scale (0 - no growth/pipette tip indent, 0.25 - pinprick growth, 0.5 - circular growth evident, 0.75 - circular growth with pinprick, 1 - dense growth throughout). The maximum score between the two bio-replicates of each isolate was used for statistical analysis, though the same statistical conclusions were obtained if the mean score was used instead. For each participant, a Wilcoxon rank sum test was used to compare vaginal and rectal isolates.

4.3.12 Data availability

FASTQ files generated for this project have been deposited at the National Center for Biotechnology Information (NCBI) Sequence Read Archive under BioProject ID PRJNA991137. All supporting tables reference are in https://github.com/microstatslab/RVVC. All phenotypic data and code required to reproduce figures, and statistical analyses are available at https://github.com/acgerstein/THRIVE_yeast-VR. Large files (e.g., BAM, VCF) are available upon request.

4.4 Results

4.4.1 THRIVE-yeast isolates in the global species phylogenies

Participants with a history of RVVC were recruited from a specialty yeast clinic in Winnipeg, Canada. Vaginal and rectal isolates were collected from swab elutes plated onto SDA and chromogenic *Candida* agar from seventeen participants who were enrolled and screened intermittently between January 2020 and November 2022. As the goal was to quantify standing genetic variation from single time point vaginal and rectal populations, we haphazardly isolated vaginal and rectal yeast isolates from each of the four participants, where we had at minimum 12 isolates from each site, for a total of 96 isolates. We refer to these isolates as the "THRIVE-yeast" isolates, following the name of our local umbrella research program that studies The Host-microbial Relationships and Immune function in different Vaginal Environments ("THRIVE," http://www.mthrive.ca). Isolates from one participant were *N. glabratus* (YST6), while *C. albicans* was isolated from the other three (TVY4, TVY10, and YST7). Three isolates (TVY10R13, TVY4R4 and YST7R13) had low depth of sequencing coverage (< 20×) and were excluded from the genomic but not phenotypic analyses. All *C. albicans* isolates were MAT-heterozygous diploids (a/α), while the *N. glabratus* isolates were all MTL1a.

The phylogenetic relationship of the THRIVE-yeast isolates was evaluated in the context of available short-read whole genome sequenced (WGS) isolates from each species. We exhaustively searched NCBI for available *N. glabratus* sequences, finding and downloading fastq data from 526 isolates (Table S4.1 at https://github.com/microstatslab/RVVC/supporting_tables). Notably, we only found a single annotated vaginal isolate and twelve stool isolates, while over 80% of the isolates are

blood isolates. We combined the fastq data from the 24 YST6 isolates and 526 global isolates to construct the largest *N. glabratus* phylogenetic tree to date. The YST6 isolates are monophyletic and cluster with 68 bloodstream isolates and three isolates of unknown provenance from the United States, Canada, and Australia (Fig. 4.1A).

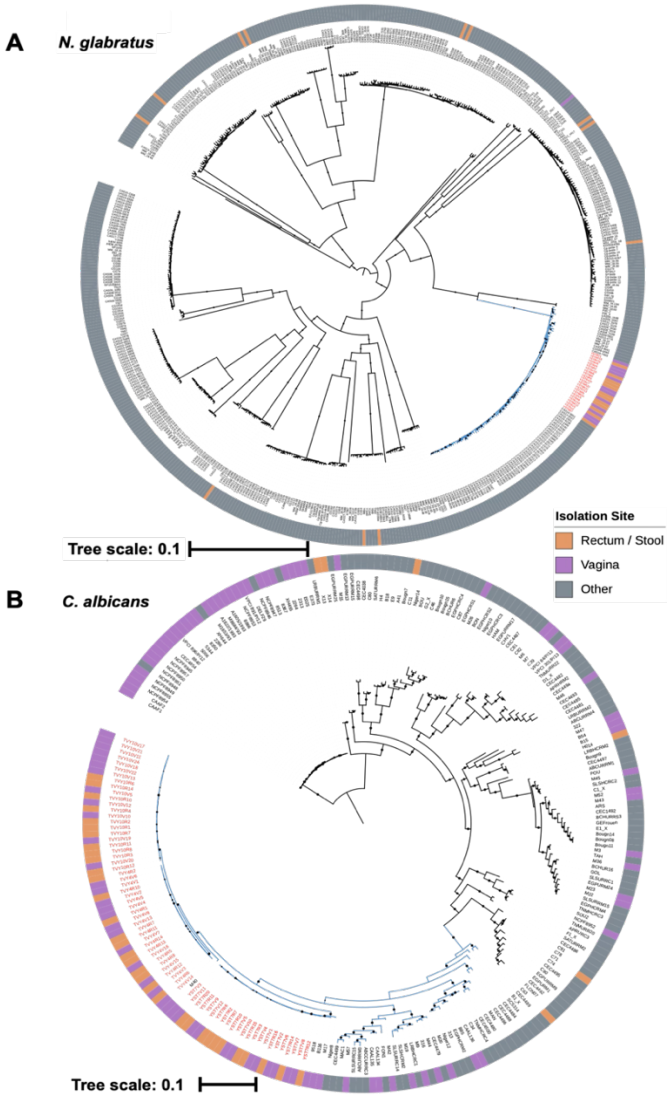


Figure 4.1: Approximate maximum likelihood phylogenies of *N. glabratus* and *C. albicans*. Approximate maximum likelihood phylogenies of (A) *N. glabratus*, including 526 global isolates, and YST6 vaginal and rectal isolates, and (B) *C. albicans*, including 182 isolates from Ropars *et al*. (2018) and vaginal and rectal isolates from TVY4, TVY10 and YST7. THRIVE-yeast isolates are indicated with red labels. The YST6 *N. glabratus* isolates are in a cluster of ST16 isolates, while the TVY4, TVY10, and YST7 *C. albicans* isolates are in clade 1 (blue branches indicate A) ST16 and B) clade 1 isolates). The *N. glabratus* phylogeny was rooted at the midpoint, and the *C. albicans* tree was rooted by *C. africana* isolates.

The widely used *C. albicans* phylogeny is composed of 182 isolates from a wide breadth of geographic and anatomical sites (Ropars et al., 2018). TVY4, TVY10 and YST7 isolates all form intra-population monophyletic groups that group in a subgroup that contains twenty-three additional isolates in clade 1, the most common clade (Fig. 4.1B). TVY4 and TVY10 isolates are beside each other and shared a common ancestry with M40, which is also a vaginal isolate from Morocco. The YST7 isolates are most closely related to three vaginal isolates (one each from Brazil, Morocco, and China) and one oral isolate from Niger. Seven of the remaining eighteen isolates in the clade 1 subgroup were also isolated from the vagina. This is a statistical enrichment for vaginal isolates compared to the rest of the isolates in clade 1 (THRIVE-yeast isolates from each participant were counted as a single isolate; Fisher exact test comparing 14 vaginal isolates out of 26 in the subgroup to 3 vaginal isolates out of 17 in the rest of clade 1, $p = 0.026$). If we discount the 35 predominantly vaginal isolates in clade 13, which is now recognized as likely a separate species (*C. africana)* (Romeo et al., 2013), clade 1 as a whole is also statistically overrepresented for vaginal isolates compared to the *C. albicans* tree in general (17 vaginal isolates in clade 1 out of 43 total isolates, compared to 18 vaginal isolates out of 107 total isolates; Fisher exact test, $p = 0.005$). Thus, although the sequenced vaginal isolates are located in six different clades, they are over-represented in clade 1 relative to a neutral expectation that vaginal isolates are equally likely to be found anywhere in the existing tree.

4.4.2 Vaginal and rectal isolates are closely related and phylogenetically overlapping.

Following the observed monophyly among isolates from the same participant, we next assessed the relatedness of vaginal and rectal isolates. We first generated phylogenies for each individual using RAxML (Stamatakis, 2014). The vaginal and rectal isolates from all four participants are phylogenetically overlapping (Figure. 4.2A). Few branches within the four phylogenies had bootstrap support exceeding 80%, yet well-supported clusters in all participants included isolates from both body sites. For each population, we also conducted a local principal component analysis (PCA) that examines the regions of high genomic heterogeneity within populations. From all participants, the regions of high differences were generally distributed throughout the genome, and the PCA failed to segregate vaginal and rectal isolates (Figure S4.1).

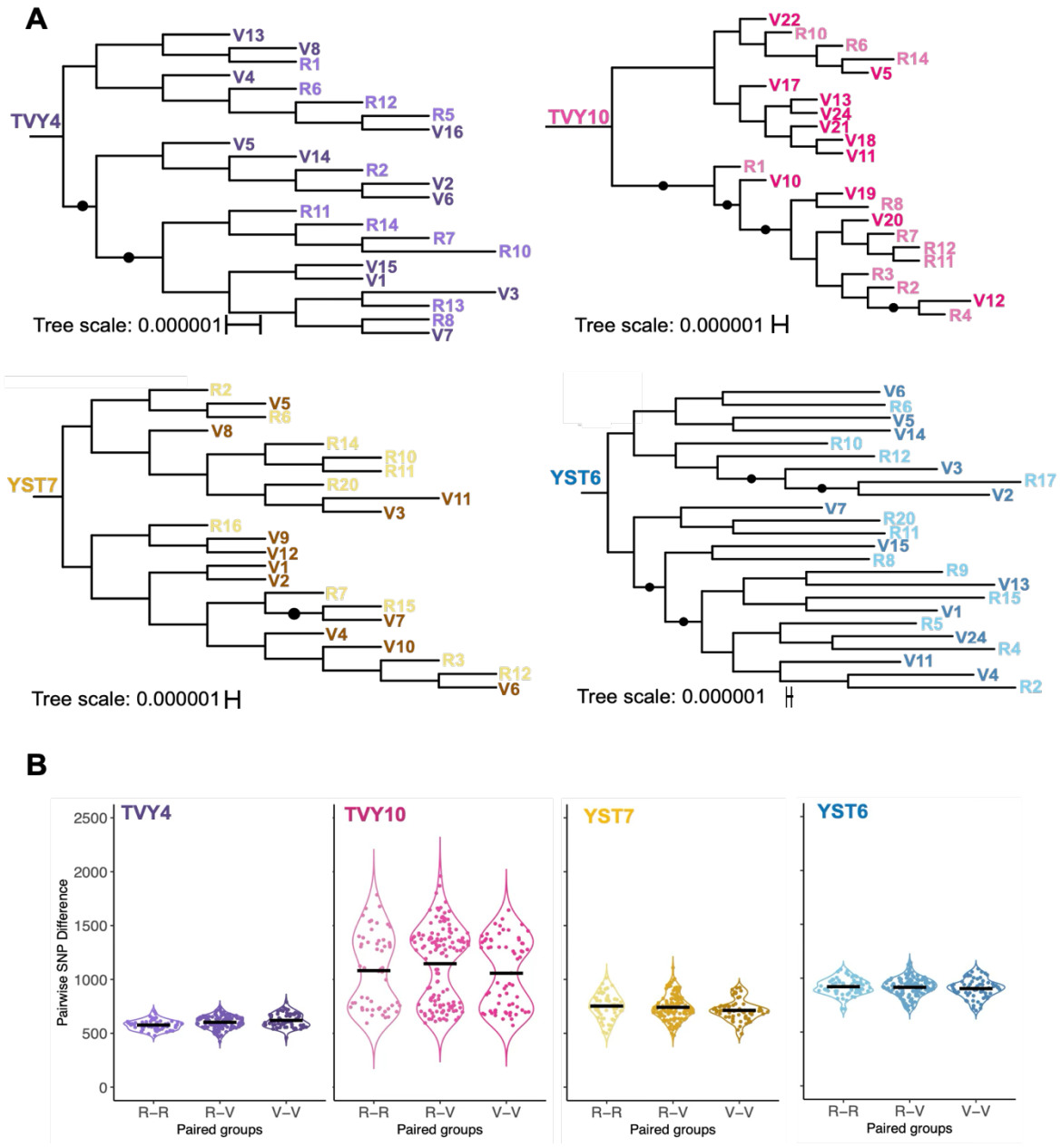


Figure 4.2: Within participant phylogenetic and single nucleotide polymorphism (SNP) analyses (A) The fine-scale phylogenetic structure among THRIVE-yeast isolates shows that vaginal and rectal isolates are closely related and do not segregate by site of isolation. Vaginal (V) and rectal (R) isolates were acquired from four participants with a history of RVVC (YST6: *N. glabratus*; TVY4, TVY10, YST7: *C. albicans*). The isolate numbers are arbitrary based on the order they were collected off culture plates. Black circles indicate branches with bootstrap support ≥ 0.8. (B) Within-participant pairwise comparison of single nucleotide positions between isolates from the indicated sites (V-V: vaginal-vaginal, R-V: rectal-vaginal, R-R: rectal-rectal) confirms that SNPs are distributed evenly among vaginal and rectal isolates.

4.4.3 Pairwise differences in single nucleotide polymorphisms among isolates per participant

Although isolates were closely related, WGS data identified SNP differences among all pairs of participant isolates. To test whether within-population vaginal diversity was lower than within-population rectal diversity, we compared pairwise SNP differences among vaginal isolates to the pairwise SNP differences among rectal isolates. To test whether there was a signal of divergence between sites, we also compared single-site differences to pairwise differences between isolates across sites. The average pairwise SNP differences between vaginal isolates were very similar to average pairwise SNP differences between rectal isolates and between isolates from different sites (Figure 4.2B). The only significant difference was in TVY4, where the average SNP differences among rectal isolates were significantly lower than the vaginal isolates (ANOVA test, YST7: $F_{2, 250} = 2.164$, $p = 0.117$; YST6: $F_{2, 250} = 0.785$, $p = 0.457$; TVY4: $F_{2, 250} = 8.599$, $p = 0.000244$; TVY10: $F_{2, 250} = 1.66$, $p = 0.192$; Tukey's HSD Test for multiple comparisons; $P_{adj} = 0.0001$). The distribution of pairwise SNP differences were fairly normal, as expected for isolates with low population structure, except that TVY10 isolates showed a bimodal distribution of pairwise SNP differences between isolates from both the vaginal and rectal sites (Figure 4.2B).

4.4.4 Minimum number of isolates for estimating standing genetic variation within participants

A major goal of our work was to compare diversity within RVVC populations to diversity observed in other contexts to make inferences about the evolutionary process based on the observed degree of standing genetic variation. However, the small number of

comparable studies sequenced different numbers of isolates, and nucleotide diversity (π) will decrease with an increased number of samples taken from a population. To quantify the scale of this effect of changing the number of isolates, we conducted a bootstrap analysis using our 12 vaginal isolates from each individual. We repeatedly resampled 3-10 isolates and recalculated diversity. For all individuals, the shape of the diversity curve with the number of isolates was very similar -an elbow was observed around n = 6 (Figure 4.3). Nucleotide Diversity in YST6 (*N. glabratus*) was two orders of magnitude lower, and there was not a consistent change in diversity with the number of isolates.

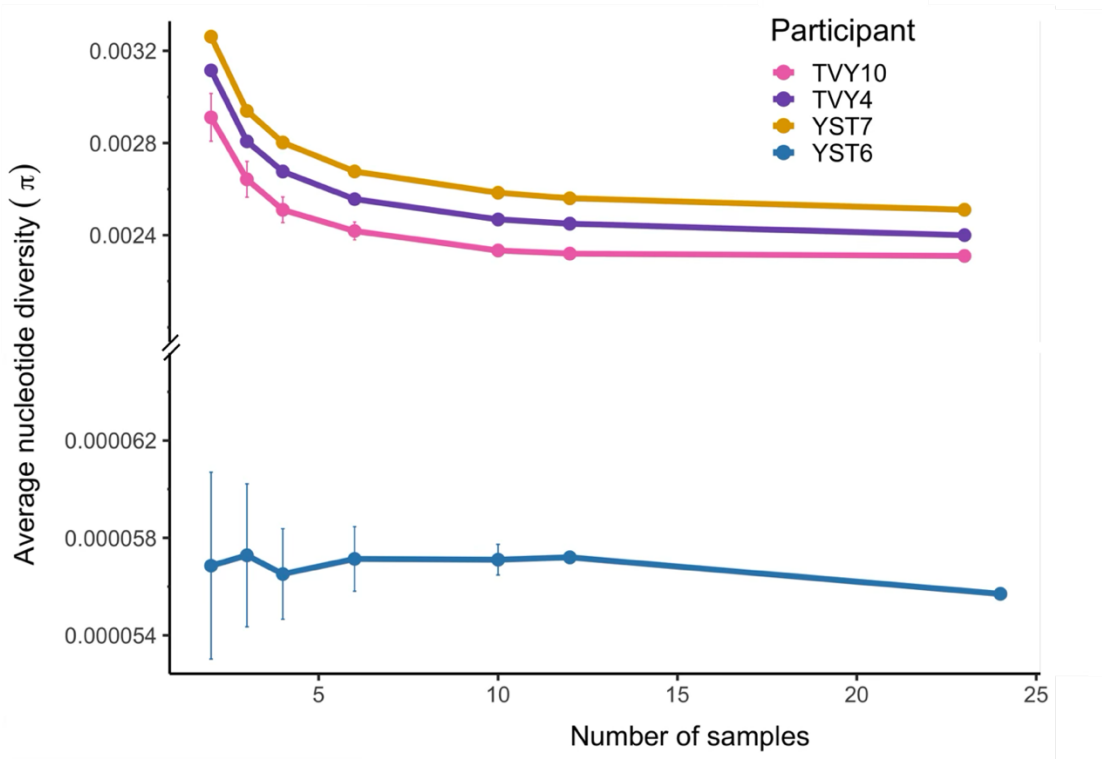


Figure 4.3: Calculated nucleotide diversity decreases then plateaus with an increasing number of samples. π was calculated for different numbers of samples from each individual. For n = 2, 3, 4, 6, or 10, the given number of samples was randomly selected from the vaginal isolate set. For n = 12, the datasets were generated by randomly choosing 6 samples from each of the rectal and vaginal isolates sets. The bootstrap analysis was done 50 times for each sample size, and the mean and standard deviation among data sets were calculated. For n = 23 (*C. albicans*, TVY10, TVY4, YST7) or 24 (*N. glabratus*, YST6), the nucleotide diversity of all samples was calculated.

4.4.5 THRIVE-yeast isolates share similar diversity as isolates from commensal and other disease settings

We downloaded the fastq files from two previous *C. albicans* studies on commensal populations (Anderson et al., 2023; Sitterlé et al., 2019) and used our pipeline to calculate the average nucleotide diversity for each. For all populations, including our own, where necessary, we down-sampled the number of isolates to three, consistent with the lowest number of isolates sampled from the commensal populations. The average nucleotide diversity was very similar across most populations (Figure 4.4A). The exception was two oral populations from two participants previously shown to have isolates from different phylogenetic clades. The YST6 vaginal isolate population was compared to fastq data from one previous *N. glabratus* study that examined 9-10 isolates from nine different blood stream infection (BSI) populations (Badrane et al., 2023). The average nucleotide diversity from YST6 vaginal and rectal populations was similar to the diversity from four participants yet much higher than the average nucleotide diversity from the other five (Fig. 4.4B). Interestingly, the four BSI populations with similar diversity to YST6 are all from an ST group (ST3), which is closely related to the cluster that contains YST6 (ST16). By contrast, the other populations are from more distantly related clades (Badrane et al., 2023). Future work in *N. glabratus* will more deeply explore whether there is a consistent relationship between clade and within-population genetic diversity.

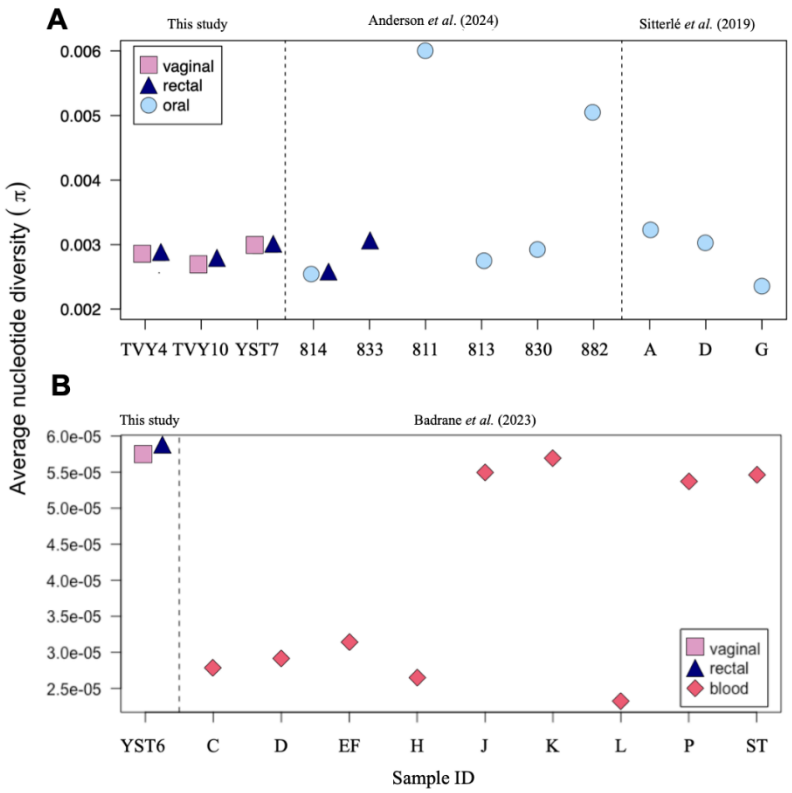


Figure 4.4: The average nucleotide diversity of vaginal and rectal populations from RVVC is similar to populations from other contexts. (A) Comparison of π in THRIVE *C. albicans* isolates from TVY4, TVY10, and YST7 to commensal isolates from two previously published studies. For accurate comparison, three randomly chosen samples were subsampled from each site in the THRIVE-yeast isolates. (B) Comparison of π in THRIVE *N. glabratus* (YST6) isolates to bloodstream infection isolates from 10 individuals in a previous study.

4.4.6 Little variation in copy number or loss of heterozygosity within populations

We examined copy number variation (CNV) and loss-of-heterozygosity (LOH) events among THRIVE-yeast isolates and their closest relatives using Y_{MAP} (Abbey et al., 2014). No CNVs were identified in any YST6 isolates (Fig. 4.5A). A single ~50 kb copy number gain on the right arm of chr3 was identified in all YST7 isolates (Figure 4.5B). This CNV is also present in the closest relative to the YST7 isolates, vaginal strain 9518, but is absent in the next two closely related strains that are also vaginal in origin (B116 and M17). As Y_{MAP} visualizations

are based on averages across 5000 bp sliding windows, we examined the region in finer detail. Coverage was measured from the BAM files to compute the depth at each position in that region. Mapped coverage was inconsistent with the profile of a typical CNV; the majority of the region had only a slightly elevated copy number relative to the rest of the genome (Fig. 4.5C). Two small (< 200 bp) regions spiked up to ~6-fold and ~14-fold coverage, the first internal to *ALS6* and the second to *ALS7* (Figure 4.5C). A third region comprised of elevated coverage maps to another gene with close homology to other genes in the genome, *CYP5*, a putative peptidyl-prolyl cis-trans isomerase (Pemberton, 2006). The region identified in Y_{MAP} is thus likely to primarily reflect an error in mapping rather than a true CNV with potential biological effects. No other CNVs or aneuploidies were identified in the other THRIVE-yeast isolates.

LOH analysis in the *C. albicans* populations was consistent with the phylogenetic analysis and diversity metrics; generally, all isolates from the same participant shared an LOH profile. The only exception was in TVY10, where an LOH tract on the left arm of chr1 was found in three rectal and 8 vaginal isolates that clustered together phylogenetically (Figure 4.5B). Some LOH regions were similar among isolates from different participants. All isolates exhibited LOH on chr3L. TVY10 and TVY4 shared a ~300 kb LOH on chr1R, and both have an LOH region on chr7R. TVY10 and YST7 shared LOH on chr5L, with distinct allelic profiles. YST7 shared LOH regions with related vaginal isolates (9518, B116, M17) but not the oral isolate Niger8 on the right arm of chr2, the left arm of chr5, and the left arm of chrR. These common LOH regions hint at the possibility of repeated selection for homozygosity in these regions.

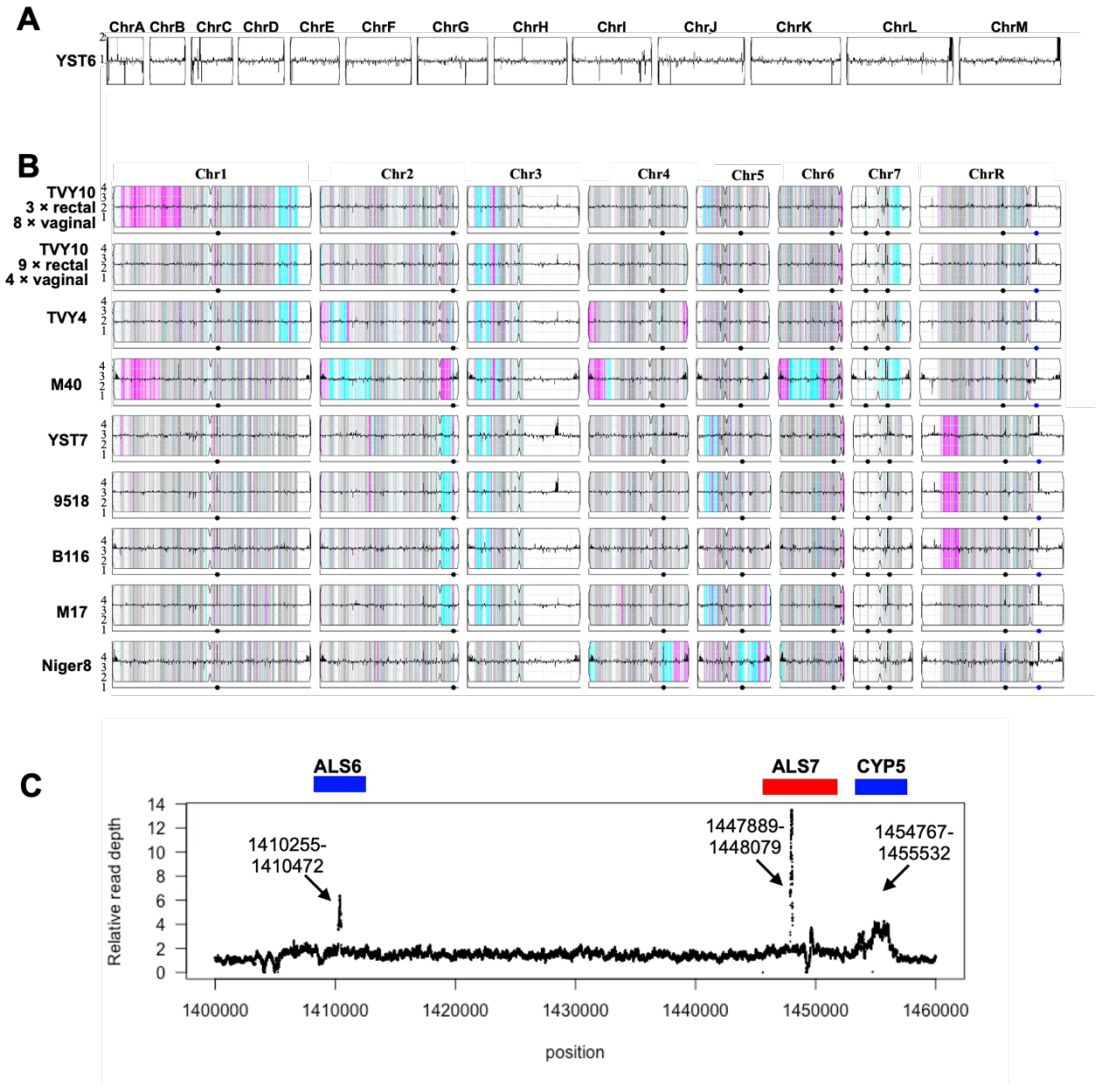


Figure 4.5: CNV and LOH profiles of THRIVE-yeast. Representative traces from (A) YST6 and (B) TVY10, TVY4, YST7, and five closely related isolates. All isolates are euploid, indicated by the horizontal black line in each panel, which indicates relative copy number by comparing the number of reads that map to each position compared to the reference genome. For *C. albicans* isolates in panel B, the density of heterozygous SNPs in 5 kb bins is shown as vertical-colored lines. Regions with heterozygous SNPs are gray, regions on homozygous SNPs are colored based on the retained SC5314 haplotype: cyan for "AA" and magenta for "BB." White indicates an ancestral LOH in SC5314. For each chromosome, the centromere is indicated by an indentation in the box. The dots on the bottom line below each box indicate the positions of major repeat sequences. (C) Fine-scale coverage mapping of the putative CNV on chr3. Shown is one representative trace from YST7R2, all isolates have a similar pattern. Gene positions above the figure are approximated.

4.4.7 Phenotypic variation

We quantified within-population phenotypic variation in parallel to genotypic variation. The average growth rate for YST6 *N. glabratus* isolates was higher than the *C. albicans* populations in both Roswell Park Memorial Institute (RPMI) medium (Fig. 4.6A) and vaginal simulative medium (VSM, Figure 4.6B). Growth rates were either the same between vaginal and rectal isolates or the rectal isolates were higher when grown in either medium (Welch Two Sample t-test, Table S4.4, rectal isolates higher in TVY4 grown in RPMI: t = 2.50, df = 16.1, p-value = 0.023; TVY4 grown in VSM: t = -2.36, df = 18.1, p-value = 0.030; YST7 grown in VSM: t = -3.63, df = 18.8, p-value = 0.002). Consistent with a visual inspection, single statistical rectal outliers with increased growth rate relative to other isolates were seen in YST7 and TVY4 grown in RPMI, and YST6 and TVY10 grown in VSM (Fig. 4.6, Rosner's test for outliers). No additional outliers were identified in the other populations.

We then compared drug resistance and drug tolerance from vaginal and rectal isolates. We conducted a pilot experiment on 24 isolates from participant YST7 to quantify variation in drug resistance for five different drugs that are indicated as treatment options by the Society of Obstetricians and Gynecologists of Canada for uncomplicated, recurrent and non-albicans VVC (van Schalkwyk et al. 2015). We also examined drug tolerance, the ability of drug-susceptible populations to grow slowly in the presence of high levels of fungistatic drugs, which also emerged as a trait that varies among different fungal species and isolates (Bhattacharjee, 2016; Salama and Gerstein, 2022). Tolerance may be implicated in the propensity to cause fungal disease in other contexts (Venkateswarlu et al., 1997) but has not previously been examined in the context of R/VVC.

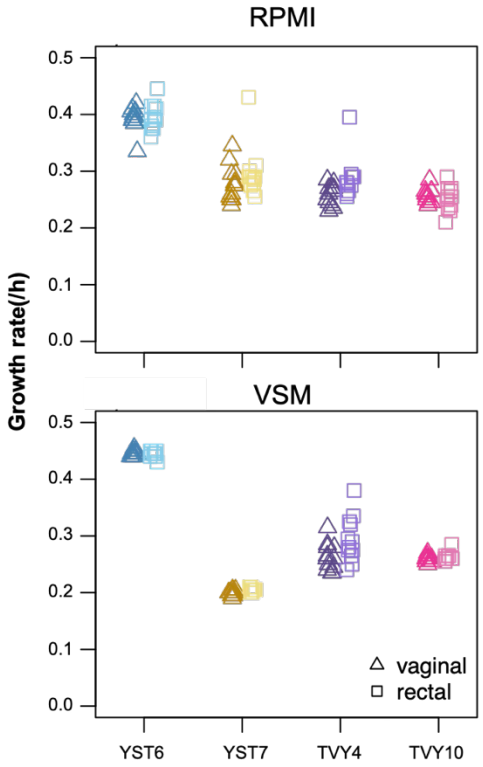


Figure 4.6: Growth rate was measured from 12 vaginal and 12 rectal isolates from each population. Optical density was recorded every 15 minutes in a plate reader with constant shaking and incubation at 37 °C. Each point represents the mean of two technical replicates for each of the two biological isolates, 24 isolates were measured for each group. The growth rate was calculated as the spline with the highest slope using a custom R script.

We found very little variation among the isolates for either phenotype in any drug (Figure S3.3). Given that, we proceeded with quantifying drug responses for just fluconazole and boric acid at pH 4.2, as these are drugs from different classes that are commonly prescribed in our local clinic. The site of isolation was only significant for YST6 BA resistance (vaginal isolates were slightly more tolerant than rectal isolates; t-test, t = 2.77, df = 18.2, p-value = 0.012, Table S4.4, Figure 4.7). Formal outlier statistical tests that grouped vaginal and rectal isolates were broadly consistent with the qualitative visual assessment,

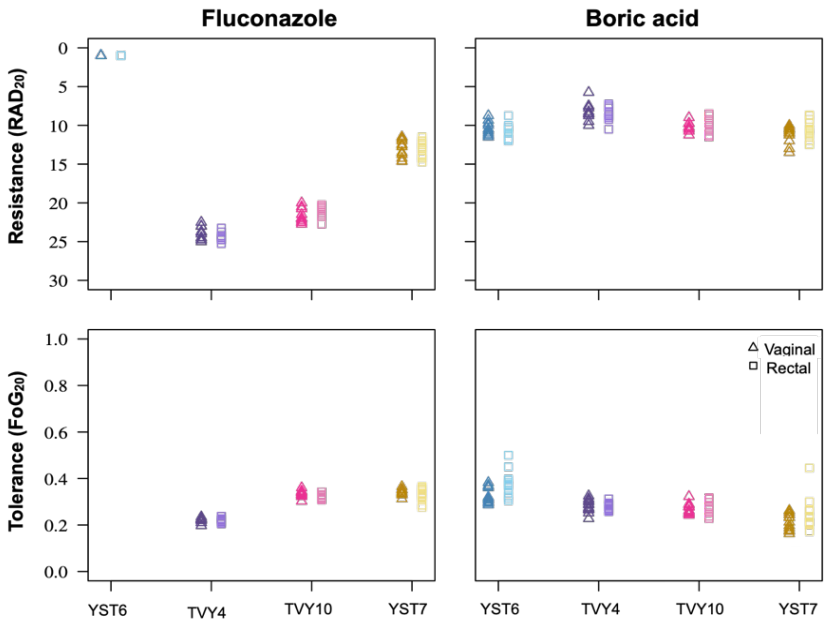


Figure 4.7: Little intra-population variation was found for drug response phenotypes measured from disk diffusion assays. Drug resistance (top panels) and drug tolerance (bottom panels) were measured for fluconazole (FLC) and boric acid (BA). Drug response was measured on pH 4 .2 Mueller-Hinton plates using the R package *diskImageR* (Gerstein et al., 2016), which computationally measures response parameters from images of disk diffusion assays. Each point represents the mean of four replicates (two technical replicates for two biological replicates).

identifying only two outlier isolates for tolerance (both rectal isolates in BA, one from YST6, one from YST). There was considerable variation among participants and isolates for invasive growth (Figure 4.8, Table S4.4 at https://github.com/microstatslab/RVVC/supporting_tables). There was no difference between YST6 or TVY10 vaginal and rectal isolates while YST7 vaginal isolates had higher invasive growth than rectal isolates ($W = 115.5$, $p = 0.0002$), and TVY4 rectal isolates exhibited higher invasive growth than vaginal isolates ($W = 372$, $p = 0.032$). The overall picture is thus that invasive growth seems to vary more between participants than between sites of isolation and that statistical differences between sites are likely due to neutral processes rather than selection for invasive growth in the vaginal environment.

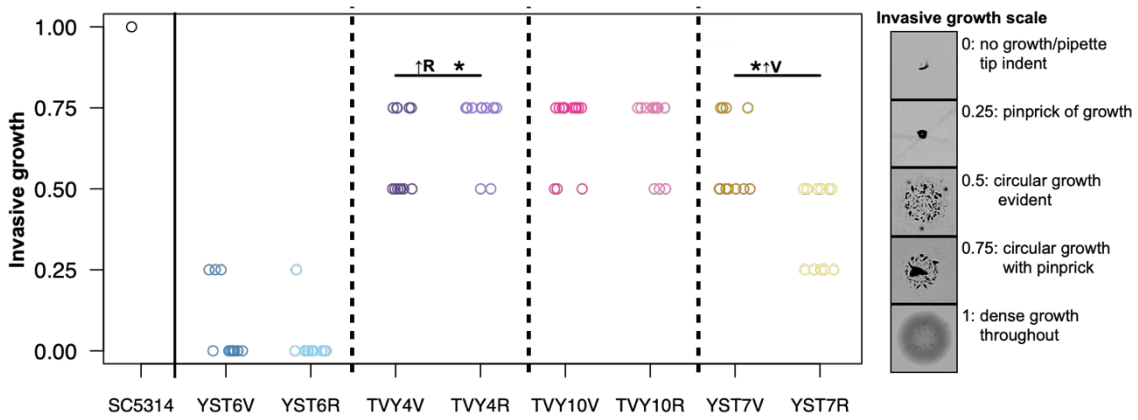


Figure 4.8: Invasive growth was qualitatively scored after growth on YPD plates for 96h. Each point indicates the maximum score between two bio-replicates for each isolate.

4.5 Discussion

The biological basis of RVVC requires further understanding. To study the vaginal yeast populations implicated in the disease, we quantified the diversity of 12 vaginal and 12 rectal isolates from each of four people with a history of RVVC who had large yeast populations at both sites at the time of sampling. In each case, the isolates formed monophyletic groups, and the vaginal and rectal isolates were phylogenetically overlapping, consistent with a common ancestral source and frequent migration between the two sites. This is in concordance with previous genetic studies that used more coarse sequencing methods, which found high genetic similarity between vaginal and rectal isolates in women with R/VVC (Araújo Paulo de Medeiros et al., 2014; Sampaio et al., 2003; Shi et al., 2007). Our phenotypic analyses are also consistent with the genetic results; there was minimal diversity in drug responses or growth ability in clinically relevant medium and no consistent difference between isolates from different isolation sites for invasive growth. Previous studies have also found that virulence factor phenotypes have similar expression among vaginal and rectal isolates in the context of R/VVC (Araújo Paulo de Medeiros et al., 2017) and among oral and rectal isolates from healthy individuals (Anderson et al., 2023).

Multiple potential evolutionary explanations are consistent with our results. It is possible that the selective pressures most influencing adaptation are similar in both environments, leading to the selection of the same traits. However, it could also be that selection cannot overcome either (or both) migration or genetic drift due to a low effective population size. The isolates come from individuals with a long history of antifungal treatment, and the yeast populations likely experience repeated bottlenecks that reduce the effective population size, thereby weakening the power of natural selection relative to

genetic drift. Nevertheless, we did observe 100s of SNP differences between all isolate pairs, indicating the presence of what has been termed microvariation (Odds et al., 2006), which is potentially sufficient for adaptation.

The observed average nucleotide diversity among populations for the RVVC isolates was higher than expected. *A priori* we predicted that genetic diversity would be highest in commensal populations from healthy individuals at sites that do not have obviously strong selective pressures acting on them and lowest in bloodstream infections which are known to have a small population of circulating yeast cells. However, genetic diversity within the three RVVC *C. albicans* populations was similar to oral and rectal isolates from seven healthy individuals (though lower than two commensal oral populations previously known to have isolates from different clades) (Anderson et al., 2023; Sitterlé et al., 2019). Given the limited sampling that has been conducted, it is difficult to make inferences about evolutionary dynamics and to benchmark "high" versus "low" diversity. There may be selection in the commensal oral (Lemberg et al., 2022) and rectal environments, involving the fixation of alleles and reducing genetic diversity to a similar degree as the RVVC populations. It could also be that the population bottlenecks in RVVC (and in at least some bloodstream infections) are not as strong as anticipated. Teasing apart these explanations will require intra-population data on many more populations from different contexts.

Importantly, we chose isolates for sequencing blind to phenotypic data, to provide an unbiased estimate of genetic diversity. Although it is tempting to pick the most diverse isolates for in-depth genomic characterization, this makes it more difficult to compare results among studies (Qu et al., 2020). The level of genetic diversity we uncovered in the vaginal yeast populations suggests a relatively high level of standing genetic variation,

particularly since the available comparable WGS studies in other contexts purposefully selected isolates that maximized phenotypical differences (i.e., (Anderson et al., 2023; Badrane et al., 2023)). Interestingly, the presence of vaginal genetic diversity is consistent with one of the earliest genetic studies, which used DNA fingerprinting to examine diversity at a single time point in up to 14 vaginal isolates from six RVVC populations (Lockhart et al., 1995).

All three *C. albicans* RVVC populations were part of a subgroup in the global phylogeny, within clade 1. This subgroup is enriched for vaginal isolates compared to the entire tree and even compared to the other part of clade 1. Compared to different clades, a greater proportion of clade 1 isolates have previously been noted to be significantly associated with superficial infections (Odds et al., 2007), including in the context of R/VVC (Ge et al., 2012; Song et al., 2022; Tian et al., 2021; Zhu et al., 2022). There may be something unique to the common ancestor of the clade 1 subgroup that makes them more amenable to colonizing and invading epithelial surfaces in general and hence able to cause vaginal disease (Giblin et al., 2001; Schmid et al., 1999). Most phenotypes of potential clinical interest have previously been found to vary among isolates within the same clade, precluding clear phenotype × clade associations (e.g., (Cravener et al., 2023; MacCallum et al., 2009)). However, it may be that more fine-scale phylogenetic resolution is required to tease apart relationships; if only a subgroup of clade 1 isolates is enriched for a particular phenotype, that might not be seen if all clade 1 isolates are grouped together. Sala *et al.* recently found that VVC isolates induced greater fungal shedding from epithelial cell culture and differently stimulated epithelial signaling pathways compared to isolates from healthy women (Sala et al., 2023). This is the clearest *in vitro* assay able to differentiate VVC and healthy isolates

phenotypically, and hence, a strong target for a Genome-wide association study (GWAS) analysis to potentially pinpoint the genetic basis of this seemingly important trait. It will be of great interest in the future to determine whether there is a genotypic association between common variants in the subgroup of clade 1 isolates (including isolates from other body sites) and their interaction with vaginal epithelial cells.

Significantly less work has been done to examine *N. glabratus* in the context of R/VVC compared to *C. albicans*, despite the increasing incidence of *N. glabratus* globally as an etiological agent of R/VVC (Kennedy and Sobel, 2010; Makanjuola et al., 2018). We found only a single vaginal isolate out of 526 total isolates with WGS data on NCBI. This sharply contrasts with *C. albicans*, where 35% of all sequenced isolates in the current phylogeny (which was updated in this study) have been annotated as vaginal origin (Ropars et al., 2018). Surprisingly, vaginal isolates form over 10% of the isolates in the *N. glabratus* MLST database (Dodgson et al., 2003). Those isolates are widely distributed among ST groups. This highlights the gap in inclusion of vaginal isolates in *N. glabratus* studies that use WGS.

4.6 Conclusion

We have conducted the most extensive study to date that employed whole genome sequencing, modern methods for calculating and comparing diversity, and high-throughput phenotypic analyses to compare vaginal and rectal isolates from participants with a history of symptomatic RVVC. We find no evidence that rectal isolates are different than vaginal isolates, which is inconsistent with the hypothesis that the GI tract is a source population for vaginal reinfection. We observed a near-identical average nucleotide diversity between our populations and some populations from commensal (*C. albicans*) and BSI (*N. glabratus*)

settings. It remains unknown whether these values are low or high relative to other body sites and commensal or infection contexts. This emphasizes the need for further investigation into diversity within fungal microbial communities across various contexts.

Chapter 5: Discussion and Conclusion

Fungal infections impose a global health burden. Among the most prominent etiologies are the opportunistic pathogens; *C. albicans* and *N. glabratus*. *C. albicans* has long been the primary cause of candidiasis while *N. glabratus* is emerging as a major threat due to its increasing resistance to antifungal drugs (Brown et al., 2012; Healey et al., 2016; Lass-Flörl et al., 2018). Both species are included on the WHO fungal priority pathogens list as fungi that pose the greatest public health threat and/or have the greatest gaps in knowledge (WHO, 2022). Investigating their population structure is key to understanding patterns of genetic variation, clonal expansion, and by extension the adaptations that drive their persistence and emergence as pathogens in human hosts.

In this thesis, I undertook a multi-faceted genomic analysis of *C. albicans* and *N. glabratus*, focusing on their global population structure, genomic variability, and intra-host diversity, with special attention to isolates associated with recurrent vulvovaginal candidiasis. By integrating whole-genome sequencing data from globally diverse isolates, alongside deep sampling from selects individuals, I addressed key questions about the phylogenetic and intra-host variation. This general discussion brings together insights from the three core chapters, expanding and reflecting on the analysis challenges that were overcome, broader implications of the results, limitations, and future directions.

This thesis marks a significant advancement in fungal population genomics through the construction of the most extensive phylogenetic framework to date for both *C. albicans* (Chapter 2) and *N. glabratus* (Chapter 3). I leveraged a whole-genome sequencing (WGS) approach across a large isolate collection and for the first time in either species, used an objective statistical framework for clade assignments. I discovered novel clades in both

species. In *C. albicans*, my analyses underscore the influence of heterozygosity on phylogenetic inference. In *N. glabratus*, I uncovered novel admixture between clusters, contributing to a growing body of literature challenging prior assumptions of strict clonality. These insights deepen our understanding of pathogen evolution and underscore the value of broad-scale comparative phylogenetics.

Before inferring the phylogenetic structure of *C. albicans* in Chapter 2, I first evaluated the impact of haplotype correction using the HaploTypo pipeline. Organisms with base ploidies above one (such as *C. albicans*) present a unique challenge in phylogenetic studies due to heterozygosity, where different alleles are present at the same genomic loci within a single isolate genome. In phylogenetics, heterozygous sites within a species have commonly been addressed in one of three ways: by using IUPAC ambiguity codes, i.e., single-letter codes that represent multiple possible nucleotides at a position (e.g., "R" for A or G, "Y" for C or T); by excluding heterozygous sites altogether, or by randomly selecting one of the two (or more) alleles (Bravo et al., 2019; Iqbal et al., 2012). This approach can generate artificial "chimeric" sequences that misrepresent evolutionary histories, especially when phase uncertainty and ambiguity coding (i.e.., using IUPAC ambiguity codes) produces haplotypes that don't exist in reality. Accurately identifying naturally occurring individual haplotypes has been shown to strengthen phylogenomic analyses using genome-wide simulated and empirical data from hummingbird (Andermann et al., 2019). Additionally, the tool HaploTypo was developed as a rigorous variant-calling pipeline for phased genomes and was validated using *Candida albicans*, yet there was no evaluation of improvements in phylogenetic inference by comparing with reference trees (Pegueroles et al., 2020). In Chapter 2 I demonstrated, using a dataset of 182 *C. albicans* isolates, that incorporating

haplotype information significantly influenced the resulting phylogenetic tree topology. Building on this finding, I extended the analysis to 1,178 isolates and constructed a haplotype-aware phylogeny. However, the empirical nature of the dataset precluded any definitive assessment of whether the observed changes in tree topology reflected improved accuracy, since no known or "true" species tree is available for comparison.

To address this limitation and determine how much the incorporation of haplotype information improves the phylogeny, a future study could conduct a simulation of populations mimicking the mutation rate and diversity of *C. albicans* across a range of heterozygosity levels, each coupled with a known reference phylogeny. These simulated datasets will enable the construction of haplotype-aware phylogenies that can be quantitatively compared to their respective true trees (and to a haplotype-unaware tree with heterozygous positions resolved using common methods). Such an approach would facilitate a rigorous evaluation of the accuracy and impact of incorporating haplotype information in phylogenomic inference. The use of SLiM 4 (Haller and Messer, 2023), a powerful forward-in-time population genetics simulator, combined with the Perfect Phylogeny Haplotyping method (Efros and Halperin, 2012), will make it possible to model realistic evolutionary scenarios and reconstruct haplotypes with high confidence. Together, these tools will provide a robust framework for assessing the role of haplotype resolution in improving phylogenetic inference for complex, heterozygous organisms like *C. albicans*.

Phylogenetic clades provide a framework for characterizing genetic diversity within species and for inferring the relative relationships and divergence patterns among clades. Using phased WGS data, I expanded upon prior *C. albicans* phylogeny (Ropars et al., 2018) identifying six novel clades predominantly originating from Asia. I proposed reassigning

retired MLST clade numbers (whose isolates got merged into existing or new clades) to these newly identified groups and transitioning from letter-based to numeric clade names to standardize the nomenclature while preserving historical continuity. In *N. glabratus*, I observed that most WGS-defined clusters align with MLST STs or closely related STs, leading to the proposal of a naming system that recognizes dominant STs within clusters. This mirrors clonal complex strategies in other microbial systems and facilitates cross-study comparisons without sacrificing phylogenetic resolution. These nomenclature efforts are more than administrative, as accurate and standardized clade names enable consistent linkage of genotypic groups to phenotypic traits across datasets. For example, if a clade is repeatedly associated with antifungal resistance, pathogenicity, or specific ecological niches, clear naming allows researchers to track, compare, and functionally investigate these patterns across studies and over time. In this way, robust and stable clade naming forms a foundation for studying within-species phenotypic variation.

Upon showing a clear clade delineation in *N. glabratus*, the natural next step will be to identify potential phenotypic differences among the STs. Unlike *C. albicans*, where several studies have explored phenotypic differences among isolates from different clades (and generally found limited or inconsistent patterns) such efforts are largely absent for *N. glabratus*, particularly with respect to non-drug-susceptibility traits (e.g., adhesion, biofilm formation, growth patterns, morphological features, virulence-associated traits, and other environmental stress responses) across STs. For example, while some works have demonstrated clades display certain trait biases (e.g., natural resistance to flucytosine in Clade 1 (Dodgson et al., 2004; Pujol et al., 2004) and acid phosphatase activity and growth in 2M NaCl between clades (MacCallum et al., 2009)), most phenotypes remain heterogeneous

both within and between clades (Hirakawa et al., 2015; Li et al., 2003). Identifying whether there are phenotypic differences among *N. glabratus* STs is particularly important, especially given the deep phylogenetic splits between clades. These levels of divergence raise the possibility that *N. glabratus* may represent a species complex, in which more pronounced clade-level differences in phenotype or clinical behavior could be expected. For example, ST7 and ST10 isolates were characterized by increased resistance to compared to other ST isolates, while isolates from those ST7 and ST15 are more likely to cause death (Zhang et al., 2024). Additionally, comparing highly admixed strains to more genetically coherent clades may help uncover lineage-specific traits or adaptations. In summary, non-antifungal-related phenotypes remain unexplored across different *N. glabratus* sequence types, representing a substantial research gap.

Despite the depth and breadth of these datasets, substantial gaps remain in the geographic and ecological representation of isolates available from both species. This is especially important because I showed that clade distribution varied by geography and isolation source, though I also showed these factors are confounded, making it hard to determine whether geographic origin, ecological niche, or both are driving observed patterns of relatedness. These limitations restrict broader conclusions regarding the global distribution of clades, the ecological specificity of clades, and the historical processes (such as migration, local adaptation, or niche differentiation) that may have shaped the evolution of both fungal populations. In *C. albicans*, my findings (and prior studies) underscore a clear underrepresentation of isolates from Africa, Australia, the Middle East, and South America. As of mid-2025, the pubMLST database (which contains 5,834 total isolates) includes only 197 African and 88 Oceanian isolates. This reveals critical regional deficiencies that

constrain global phylogeographic reconstruction and may obscure region-specific evolutionary trajectories. My usage of an expanded sampling dataset, which incorporates significantly more isolates from Asia, contrasts with the phylogeny by Ropars et al. (Ropars et al., 2018) that underrepresented Asian isolates. This broader representation revealed multiple predominantly Asia-specific clades that were absent in previous analyses, underscoring how incomplete regional sampling can obscure detection of truly distinct populations. In *N. glabratus*, sampling is heavily skewed toward isolates from Europe and North America.

There is also significant bias in both species' isolate sets towards isolation site. In *N. glabratus* the isolates predominantly derived from bloodstream infections (an otherwise sterile anatomical site). Vaginal, gastrointestinal, and environmental isolates remain scarce. and efforts should include both commensal and disease-associated contexts to better reflect its full biological spectrum. A similar site-specific bias also exists in *C. albicans*, where sampling has largely focused on clinical isolates. My inclusion of eight environmental isolates from the United States reveals that while most cluster with human-associated isolates, at least one falls at the extreme edge of diversity between *C. albicans* and *C. africana*. This suggests that unsampled environmental niches could harbor substantial, previously unrecognized diversity. To address these limitations, broader and more representative sampling strategies are urgently needed, and it would be interested to phenotypically compare the environmental isolates to clinical ones. Moreover, deliberate and ethically grounded sampling strategies could consider whether there are isolate differences between populations that limit infection control to ethnobotanical treatments compared to those with access to approved pharmaceutical regimens. It would be interesting to test whether these

individuals harbor fungal populations that are less influenced by clinical selective pressures, potentially offering insight into local baseline population structures. Incorporating this ecological dimension is especially important for detecting geographical signatures and understanding the historical and environmental factors that have shaped prevalence of certain clades.

In both *C. albicans* and *N. glabratus*, I observed that karyotypic genomic variation (such as aneuploidy and CNVs) was more frequent in *C. albicans* (~9.5%) than in *N. glabratus* (4%). However, these alterations did not show strong associations with geography or anatomical source. This contrasts with previous findings in *C. albicans*, where specific aneuploidies appeared linked to particular niches; for instance, trisomies of chr6 were frequently observed in isolates from the oral cavity (Forche et al., 2018), while chr7 trisomies arose during colonization of the murine gastrointestinal tract, where they conferred increased fitness (Ene et al., 2018). Reduced susceptibility of *C. albicans* to echinocandin have been linked to aneuploidies of Chr5 and Chr2 (Sah et al., 2021). CNVs have also been shown to be expandable and reversible serving as drivers of rapid adaptation to antifungal drugs (Todd and Selmecki, 2020). In *N. glabratus,* prevalence of chrE, a chromosome with relevance to azole susceptibility, was observed in Chapter 3. These patterns support the hypothesis that structural variation in these species is likely transient and context-dependent, arising in response to local pressures such as antifungal treatment rather than representing a stable genome change. To further probe the ephemeral nature of these aneuploidies, I compared genome-wide heterozygosity levels in aneuploidy to non-aneuploid chromosomes in *N. glabratus* and found no association of loss of heterozygosity with aneuploid chromosomes; supporting the view that such events are short-lived and

possibly purged over time. A compelling future direction would be to assess heterozygosity specifically on aneuploid chromosomes in *C. albicans*, especially triploid ones, to better understand if and why specific haplotypes are preferred for duplication.

In Chapter 3, I showed that *N. glabratus* WGS clusters are deeply diverged phylogenetically, suggests that *N. glabratus* may in fact represent a species complex rather than a single cohesive species. *Nakaseomyces bracarensis* and *Nakaseomyces nivariensis* have been previously classified as separate species within the *Nakaseomyces* clade. Specifically, *N. bracarensis* was characterized based on PCR fingerprinting and sequence divergence in the ribosomal D1/D2 domains (Correia et al., 2006), while *N. nivariensis* was delineated using ITS region sequence analysis (Borman et al., 2008). However, the *N. bracarensis* isolate I intended to use as an outgroup clustered with *N. glabratus* ST16. Whether this is true of other *N. bracarensis* isolates and how *N. nivariensis* isolates are related to *N. glabratus* needs to be determined.

One of the key contributions of my thesis was the analysis of intra-host diversity in individuals with RVVC. My analyses of 12 isolates each of vaginal and rectal isolates from the same individuals revealed high genetic similarity and phylogenetic overlap between anatomical sites. This supports a model of frequent migration between rectal and vaginal populations. Importantly, although some genetic microvariation was detected (hundreds of SNPs between isolates), we found limited phenotypic diversity, suggesting that selection may be weak, overridden by migration, or acting on traits not captured by *in vitro* assays. Interestingly, the nucleotide diversity within these RVVC populations was comparable to that of oral commensal populations in healthy individuals, challenging assumptions that clinical infections always represent genetic bottlenecks. This observation is particularly

notable given that RVVC is typically treated with topical antifungals, which would not be expected to impact gastrointestinal or oral microbial populations. This suggests that population structure in RVVC may be more dynamic and diverse than previously appreciated. Furthermore, our identification of a vaginally enriched subgroup within *C. albicans* clade 1 highlights the potential for fine-scale genotype-phenotype associations and warrants deeper investigation.

Building on the findings of this thesis, several clear directions emerge for extending the analysis of intra-host diversity in individuals with RVVC. One immediate next step is longitudinal sampling of isolates from the same individuals over time, particularly across symptomatic and asymptomatic phases. This would help determine whether recurrent infections are due to reactivation of the same strain, reseeding from a persistent reservoir (such as the rectum or GI tract), or reinfection by new genotypes. Expanding sampling to additional anatomical sites (such as the oral cavity, gastrointestinal tract, or skin) could also shed light on overlooked reservoirs and migration dynamics between body sites. These approaches require coordination with clinician to manage longitudinal patient cohorts, yet they would yield valuable insights into the temporal and spatial distribution of *C. albicans* and *N. glabratus* within hosts. Beyond sampling, future work should pursue more detailed genotype-phenotype association studies. The identification of a vaginally enriched subgroup within *C. albicans* clade 1 suggests the potential for niche adaptation that warrants deeper functional exploration. Phenotypic profiling under more physiologically relevant conditions (e.g., co-culture with vaginal epithelial cells, mucosal biofilm models, or low-oxygen environments) could uncover traits not captured by standard *in vitro* assays. Additionally, integrating fungal genomic data with host genotyping and vaginal microbiome composition

may help uncover host-microbe interactions that shape fungal population structure and drive recurrence. Finally, comparative studies across different clinical contexts-such as oral or gastrointestinal candidiasis-would clarify whether the patterns of diversity and limited bottlenecking observed here are unique to RVVC or reflect a broader feature of mucosal *C. albicans* and *N. glabratus* colonization. Together, these future directions move toward a more ecologically informed understanding of *C. albicans* and *N. glabratus* persistence, adaptation, and host association in recurrent infection.

Beyond biological findings, this work highlighted enduring bioinformatics challenges in analyzing *Candida* spp. genomes. The lack of standardized pipelines for variant calling in these species compounded by repetitive regions in the genome hinders reliable variant detection. Repetitive sequences are especially problematic as they confound alignment and variant calling and often lead to false positives. To mitigate these issues, I filtered out repetitive regions prior to variant calling, which reduced false positives but likely excluded biologically meaningful variation as well. These trade-offs are important: repetitive regions may harbor adaptive changes, particularly CNVs, yet are often discarded in standard pipelines due to technical limitations. Developing a standardized variant calling protocol (whether across fungal species or even within a single species) is a relevant and complex undertaking. The challenges are reminiscent of those faced in human clinical genomics, where a national workgroup convened by the U.S. Centers for Disease Control and Prevention crafted recommendations to standardize how sequence variants are described in clinical next-generation sequencing variant files (Lubin et al., 2017). The workgroup emphasized the need for alignment to a common reference sequence, well-defined variant-caller settings, the use of precise genomic coordinates, and consistent gene and variant

naming conventions. Following these recommendations, multiple works have been published on best practices for human variant analysis (Koboldt, 2020; Krusche et al., 2019; Olson et al., 2023). A similar community effort is urgently needed in fungal genomics to improve cross-study consistency, reproducibility, and the biological interpretability of variant data.

Together, this thesis sets up a phylogenetic framework of the population of *C. albicans* and *N. glabratus*, shedding light on clade structure, genomic diversity, and host-associated variation. By harmonizing phylogenetic frameworks, identifying cryptic genetic diversity, and revealing gaps in current sampling, we lay a foundation for more comprehensive and representative future studies. Ultimately, integrating large-scale genomics with focused within-host and functional analyses will be crucial to decipher the evolutionary strategies and clinical implications of these important fungal pathogens.

Chapter 6: Appendix

6.1 Supplementary figures for Chapter 2

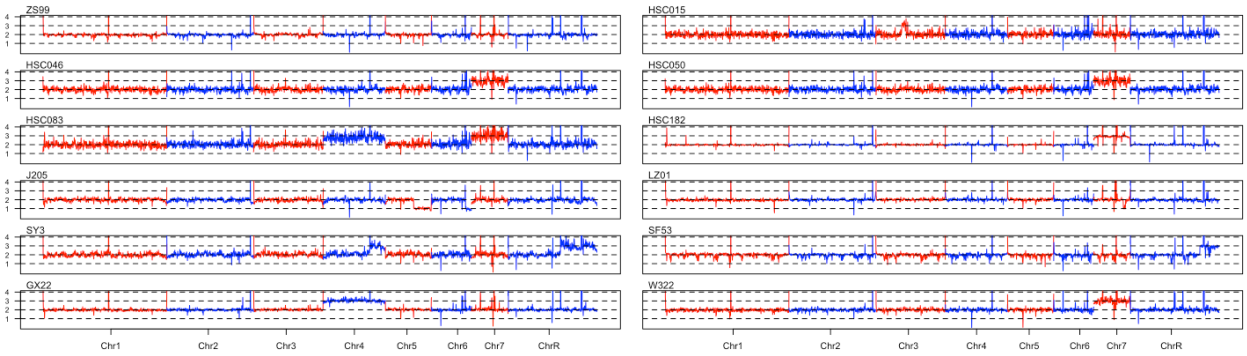


Figure S2.1: Example ploidy tracks used for quantification of aneuploidies in *C. albicans* generated with a R script. The alternating blue and red segments represent chromosomes normalized to a baseline ploidy of 2. Aneuploidies (e.g., trisomy of chromosome 7 in isolate HSC046) and copy number variations (e.g., on chromosome 3 in isolate HSC015) are shown as deviations from the baseline ploidy.

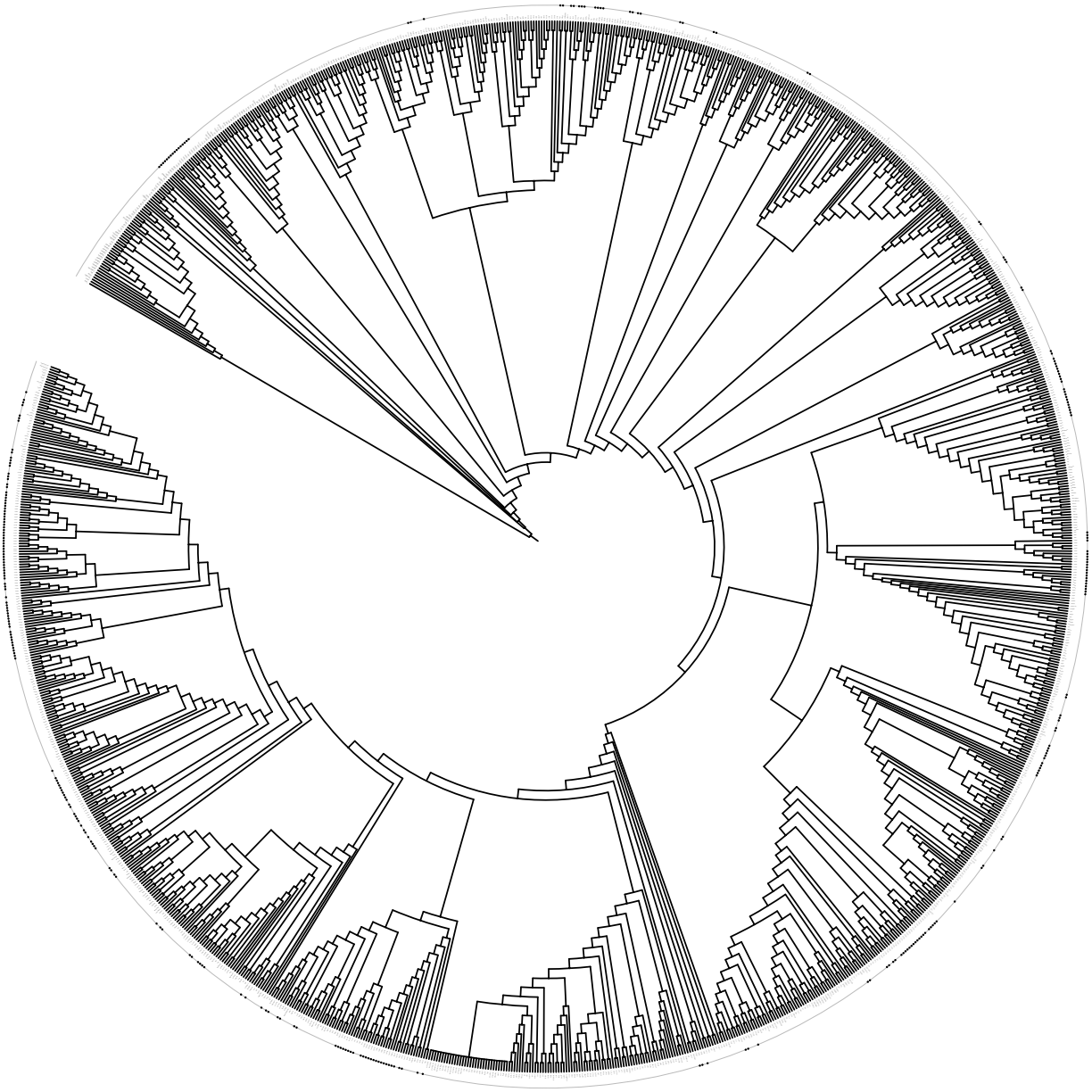


Figure S2.2: Maximum-likelihood phylogeny of *C. albicans* isolates used in this study, indicating the positions of all intrapopulation isolates as black dots in the outer ring. Variants were called using Haplotypo, and the phylogeny was constructed with FastTree and visualized in iTOL. Branch lengths were ignored to improve the clarity of isolate placement.

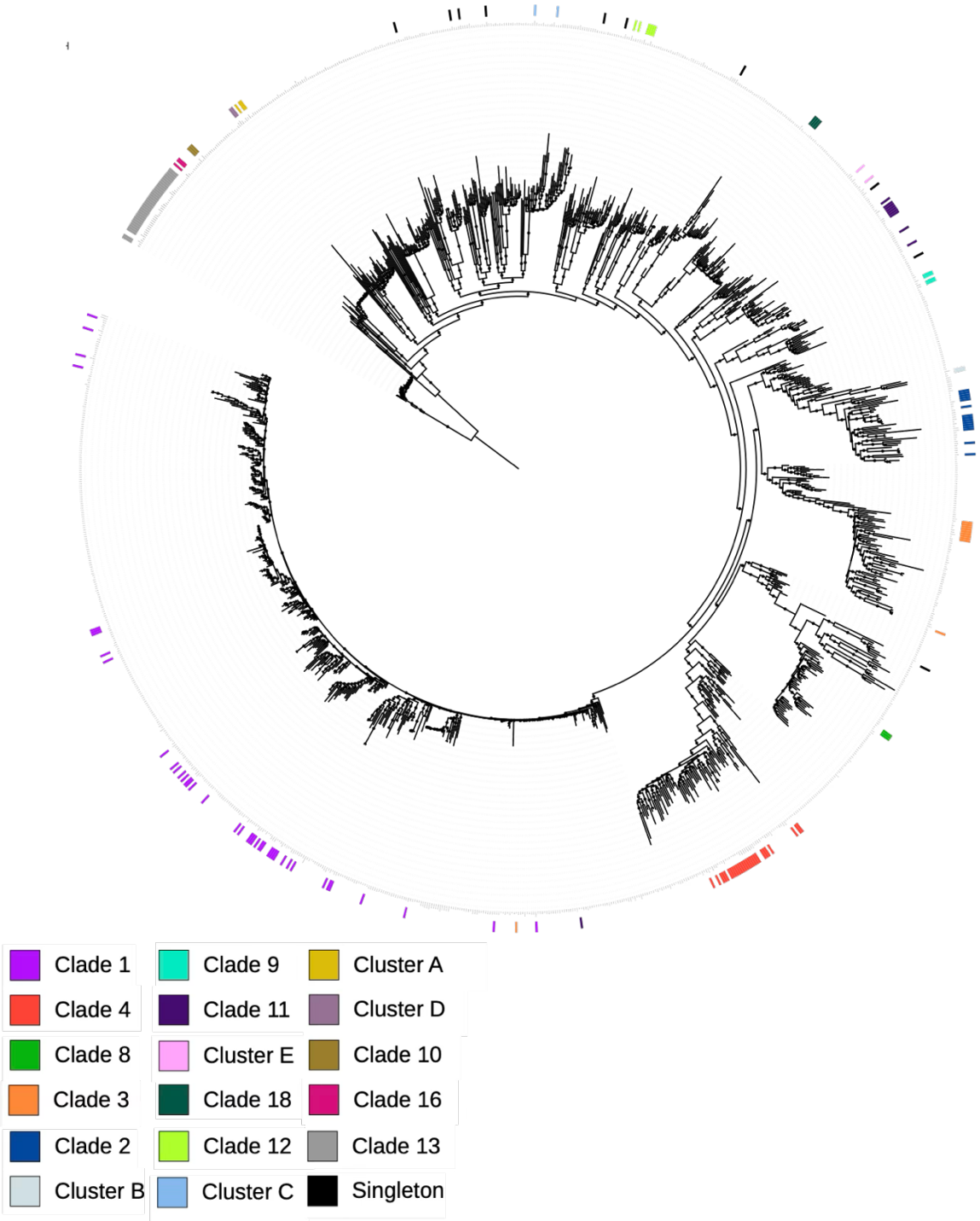


Figure S2.3: Maximum-likelihood phylogeny of *Candida albicans* isolates. The outer ring represents the position of isolates in Ropars *et al.* (2018) and the colors represent the clades assigned to these isolates by Ropars *et al.* (2018). The phylogeny was rooted with *C. africana* (clade 13)

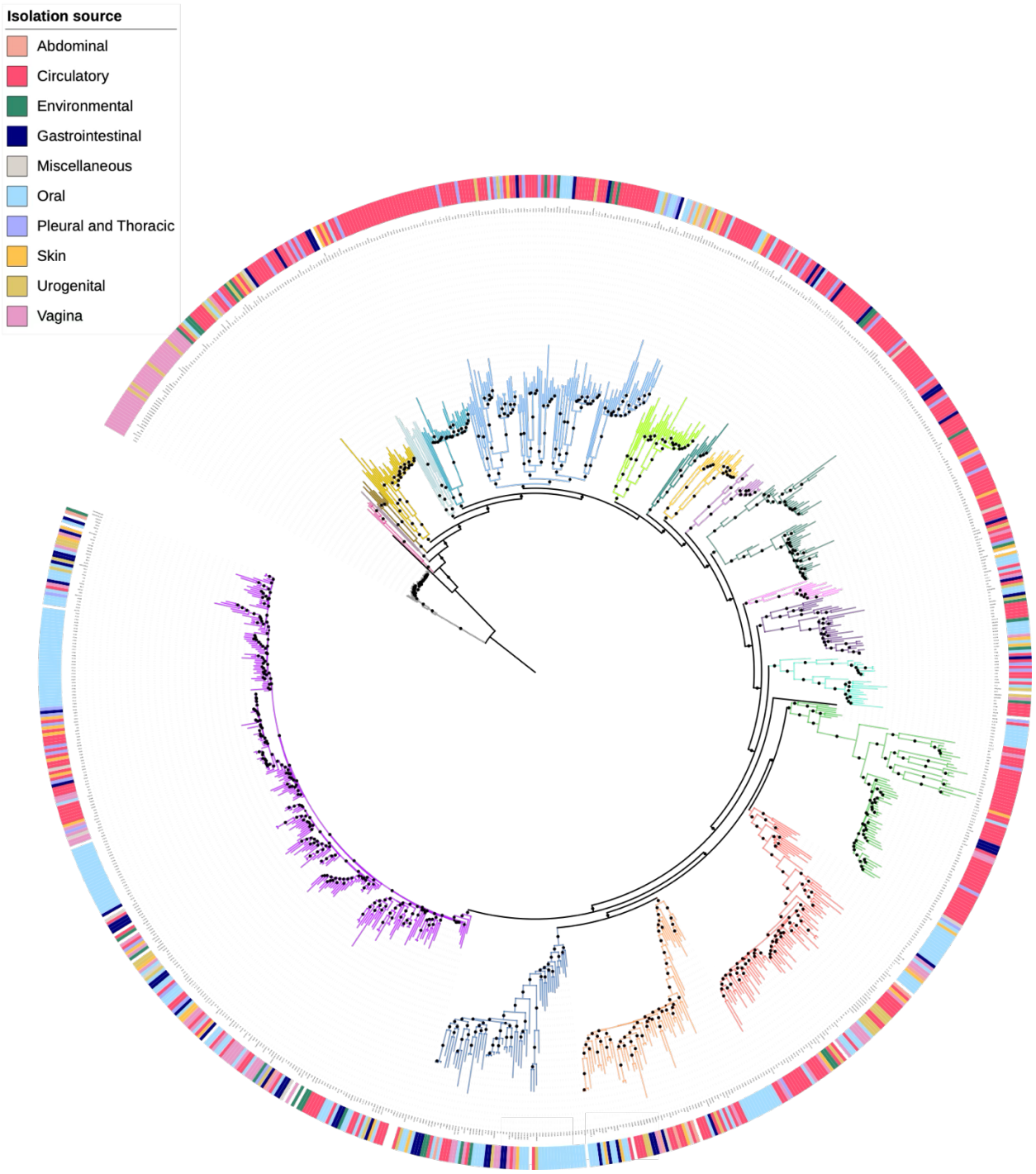


Figure S2.4: Maximum-likelihood phylogeny of isolates indicating the source of isolation. Branches are colored according to the clades to which the isolates belong. Black circles on the branches indicate bootstrap support values greater than 0.9. The outer ring represents the source of isolation, with colors corresponding to different sources.

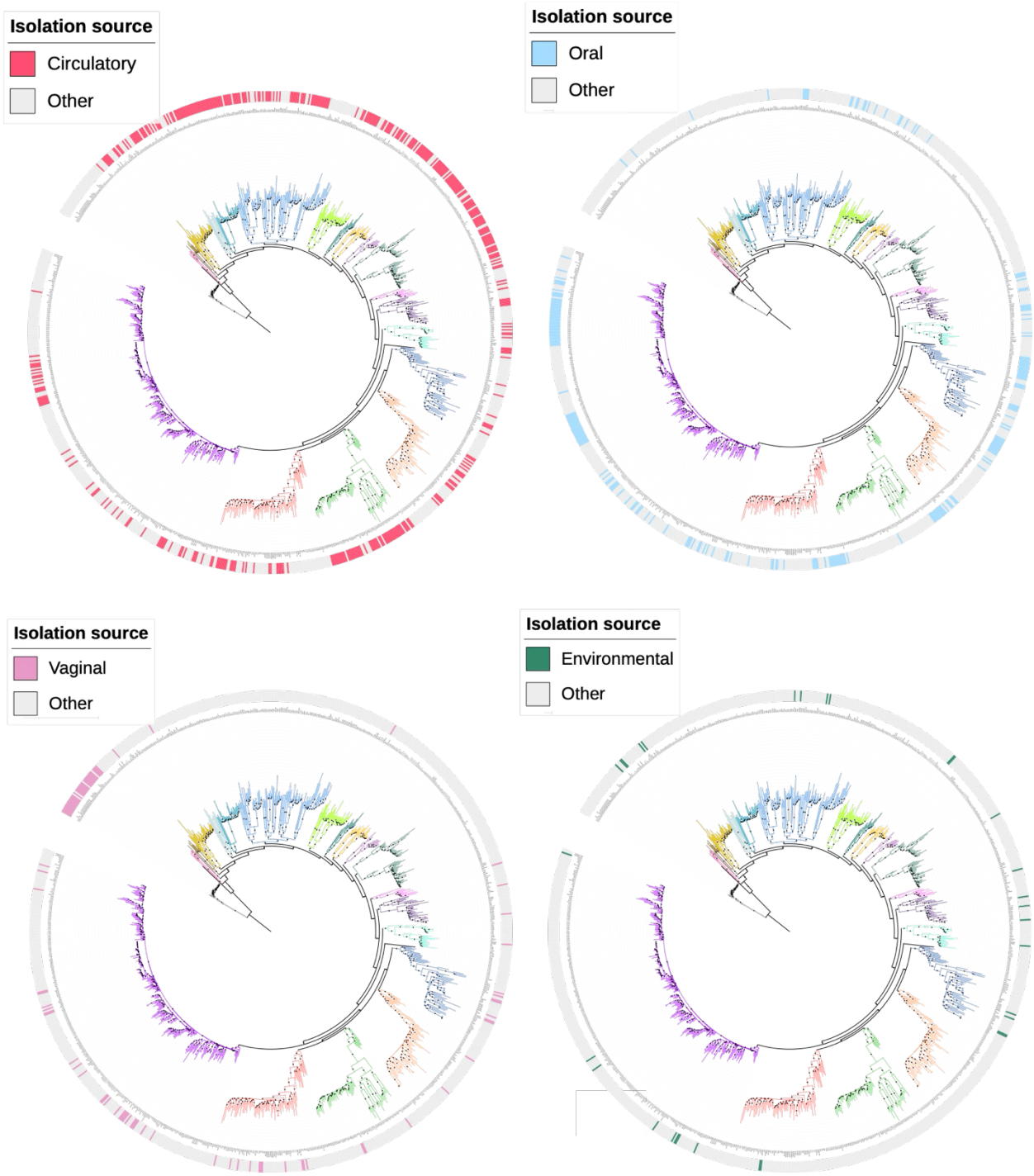


Figure S2.5: Maximum-likelihood phylogenies depicting the distribution of the *C. albicans* isolates by the source of isolation. Branches are colored according to the clades to which the isolates belong. Black circles on the branches indicate bootstrap support values greater than 0.9. The outer ring represents the source of isolation, with colors corresponding to different sources. Four main isolation sources have been shown due to preponderance and/interest.

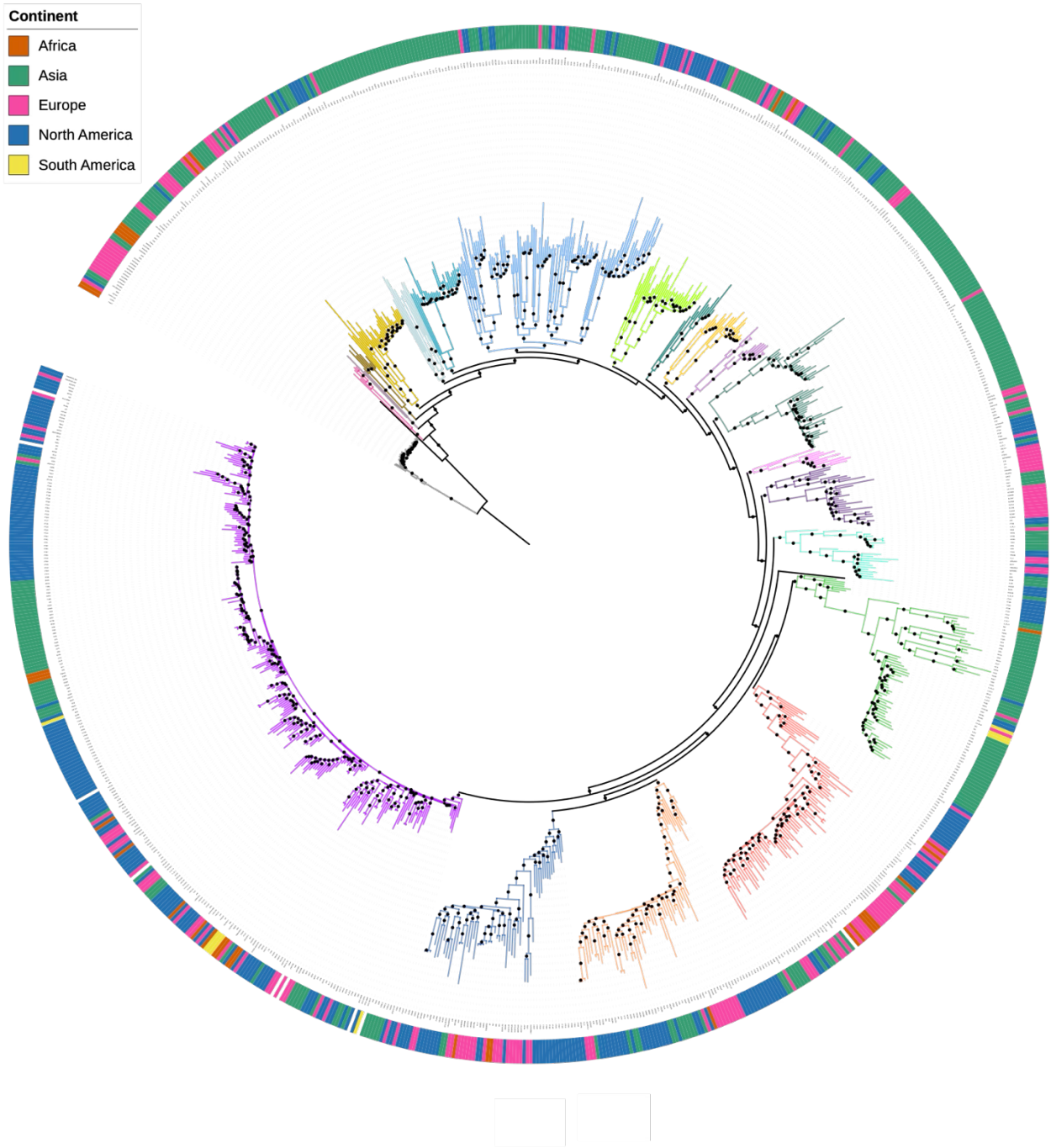


Figure S2.6: Maximum-likelihood phylogeny of *C. albicans* isolates by geography. Branches are colored according to the clades to which the isolates belong. Black circles on the branches indicate bootstrap support values greater than 0.9. The outer ring represents the geographical source of isolation, with colors corresponding to different continents.

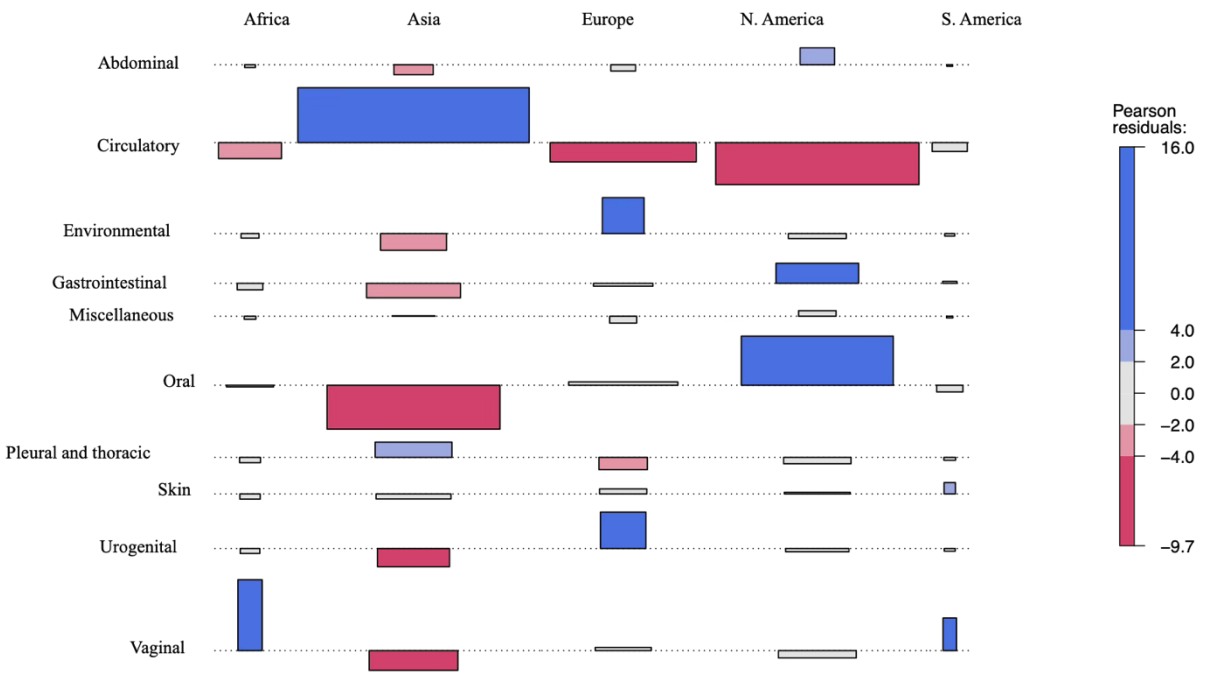


Figure S2.7: Geographic structure in isolation source and clade distribution of *C. albicans* isolates. Association plot showing deviations from independence between isolation source and continent. Each bar represents a cell in the contingency table, with height and color indicating the direction and strength of standardized residuals. The width of the bar corresponds to the marginal frequency. Blue bars represent combinations that occur more frequently than expected under independence; red bars indicate underrepresentation. The strong association (χ^2 $p < 0.00001$, Cramér's $V = 0.55$) indicates that isolation source is strongly nested within geography, reflecting structured sampling or ecological differences across regions.

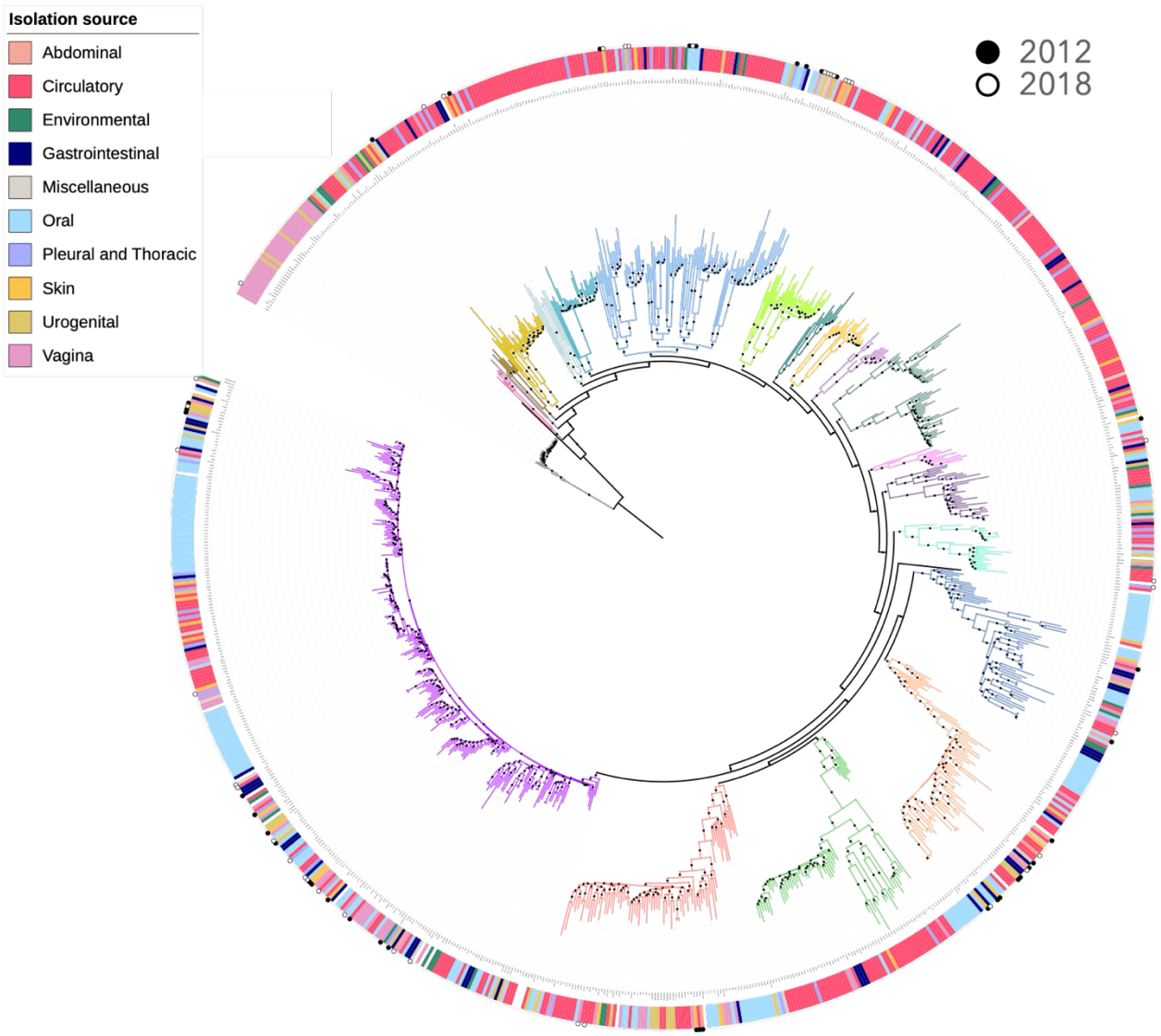


Figure S2.8: Maximum-likelihood of *C. albicans* isolates showing the phylogenetic placement of Manitoba isolates. The outer ring represents the source of isolation, with colors corresponding to different sources. Black circles on the branches indicate bootstrap support values greater than 0.9. Open and closed circles on the outer ring denote Manitoba isolates from 2018 and 2012, respectively.

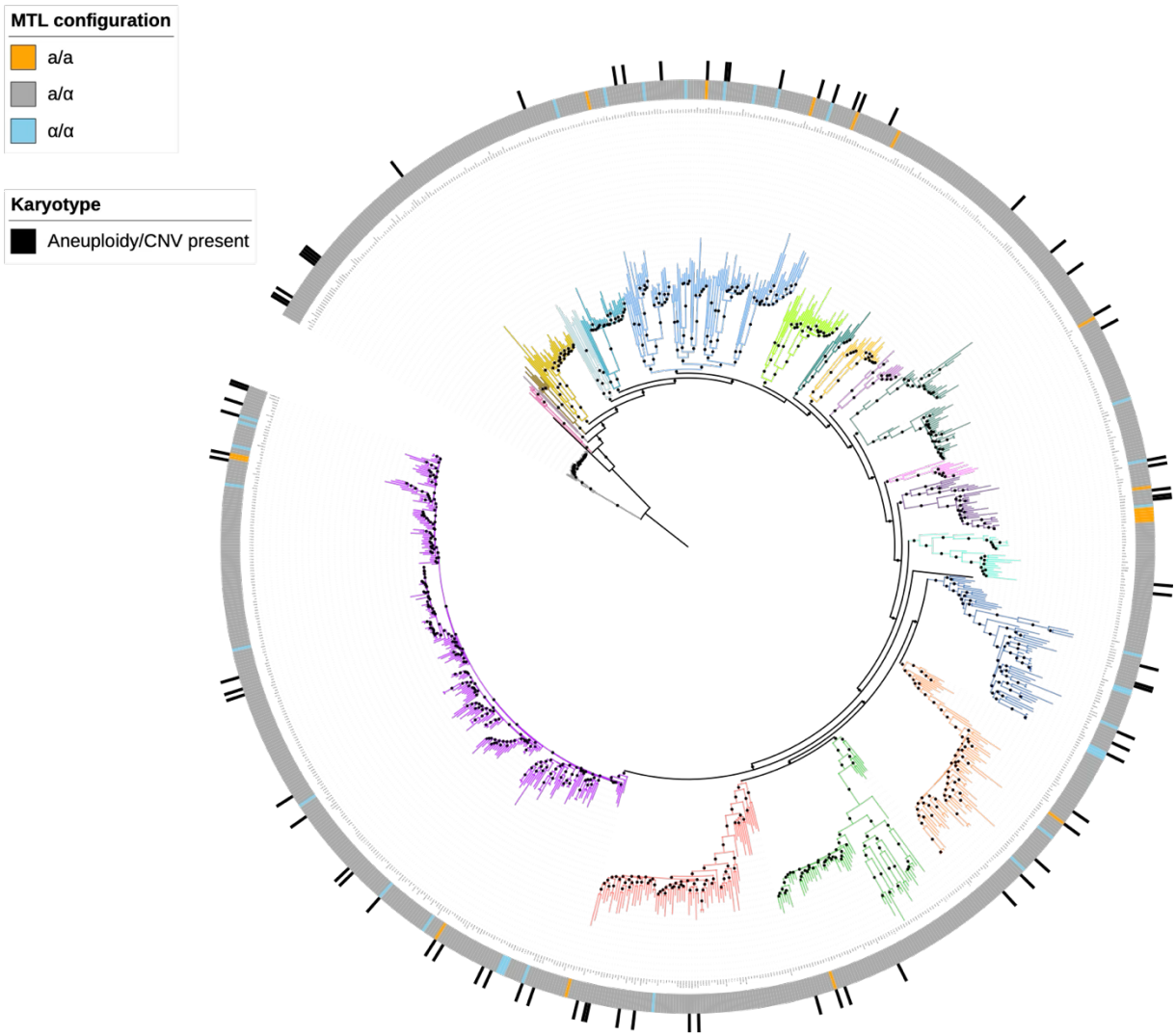


Figure S2.9: Distribution of aneuploidies, CNVs, and MTL locus configurations of *C. albicans* isolates across the phylogeny. Maximum likelihood phylogeny of all isolates, with two annotation rings. The outer ring indicates isolates exhibiting aneuploidies and/or copy number variations (CNVs), while the inner ring represents the distribution of different MTL locus configurations.

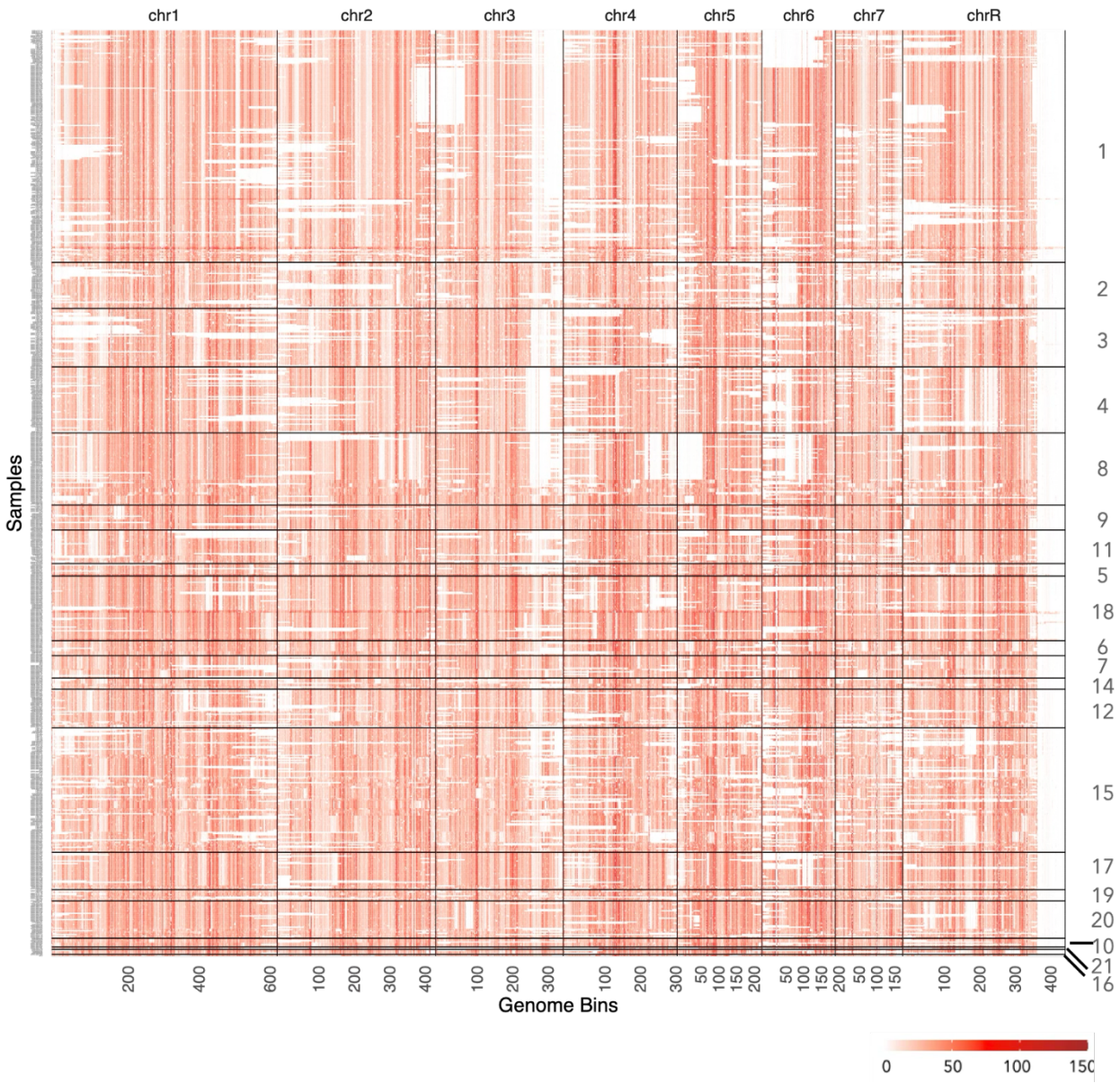


Figure S2.10: Density of heterozygous SNPs in 744 *C. albicans* isolates, in 5 kb windows. Each row represents an isolate. Isolates are ordered according to their order on the phylogeny. Thick vertical black lines delimit chromosomes (from 1 to 7 and R). Horizontal white stripes are indicative of recent LOH events. The scale bar represents density of heterozygous SNPs per 5 kb window, from a low density(white for 0) to a high density in dark red.

6.2 Supplementary figures for Chapter 3

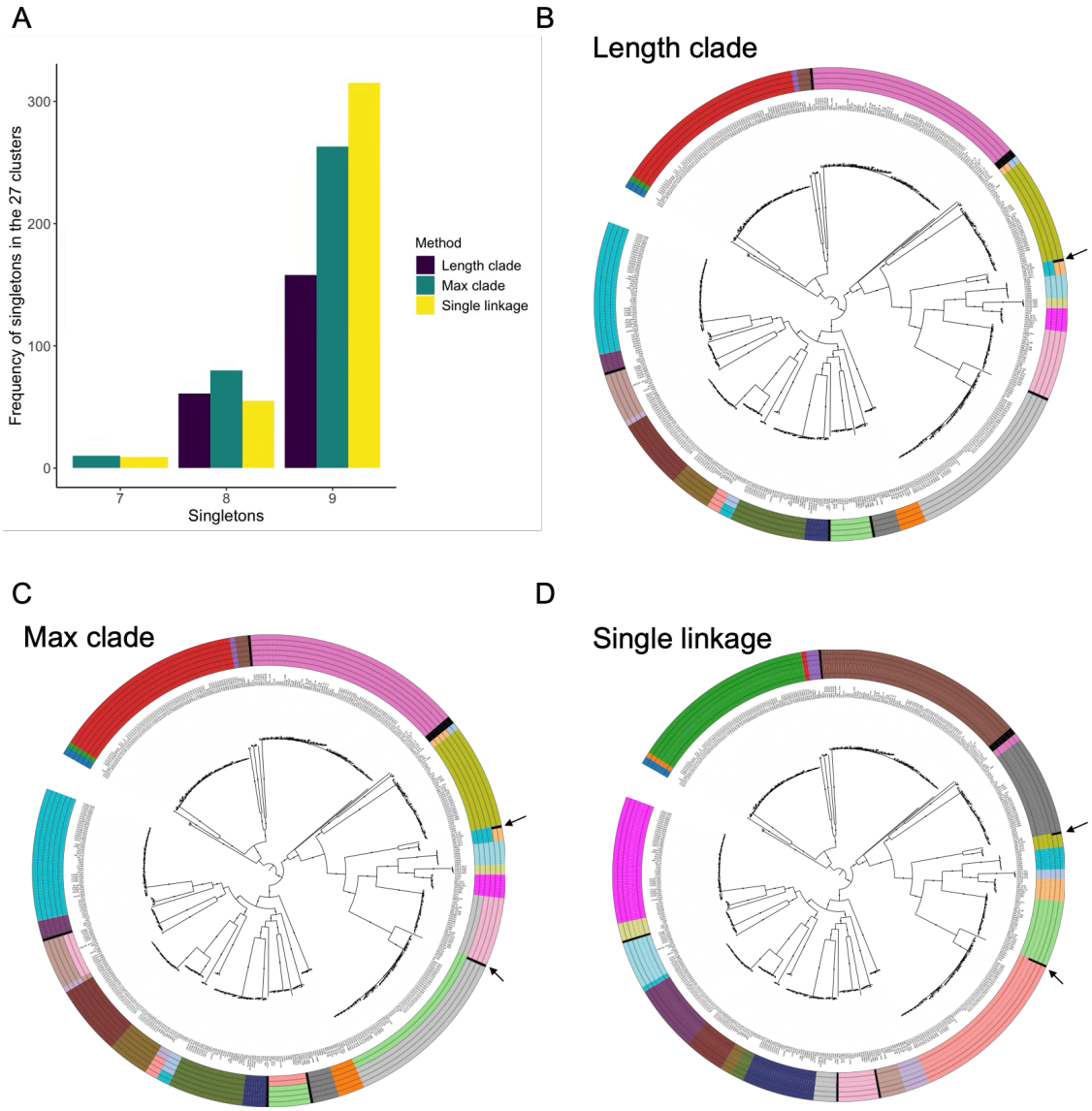


Figure S3.1: Comparison of singleton distributions within the 27 clusters predicted by three TreeCluster strategies across consecutive branch length thresholds (length clade: 0.0023–0.018; max clade: 0.0099–0.0361; single linkage: 0.00354–0.004) across the *N. glabratus* phylogeny. (A) Frequency of singletons in the predicted 27 clusters across the three TreeCluster methods. (B–D) Phylogenetic positions of isolates that contributed to variation in singleton counts under each clustering strategy. For each strategy, cluster designations corresponding to threshold boundaries with shared singleton counts are shown as colored concentric rings around the phylogeny. Length clade has four such boundaries (due to having only two singleton groups, 8 and 9, within the 27 predicted clusters), whereas max clade and single linkage each have six boundaries (due to having three singleton groups, 7, 8, and 9, within the 27 predicted clusters). Arrows indicate the two isolates responsible for deviation from the dominant 9-singleton frequency.

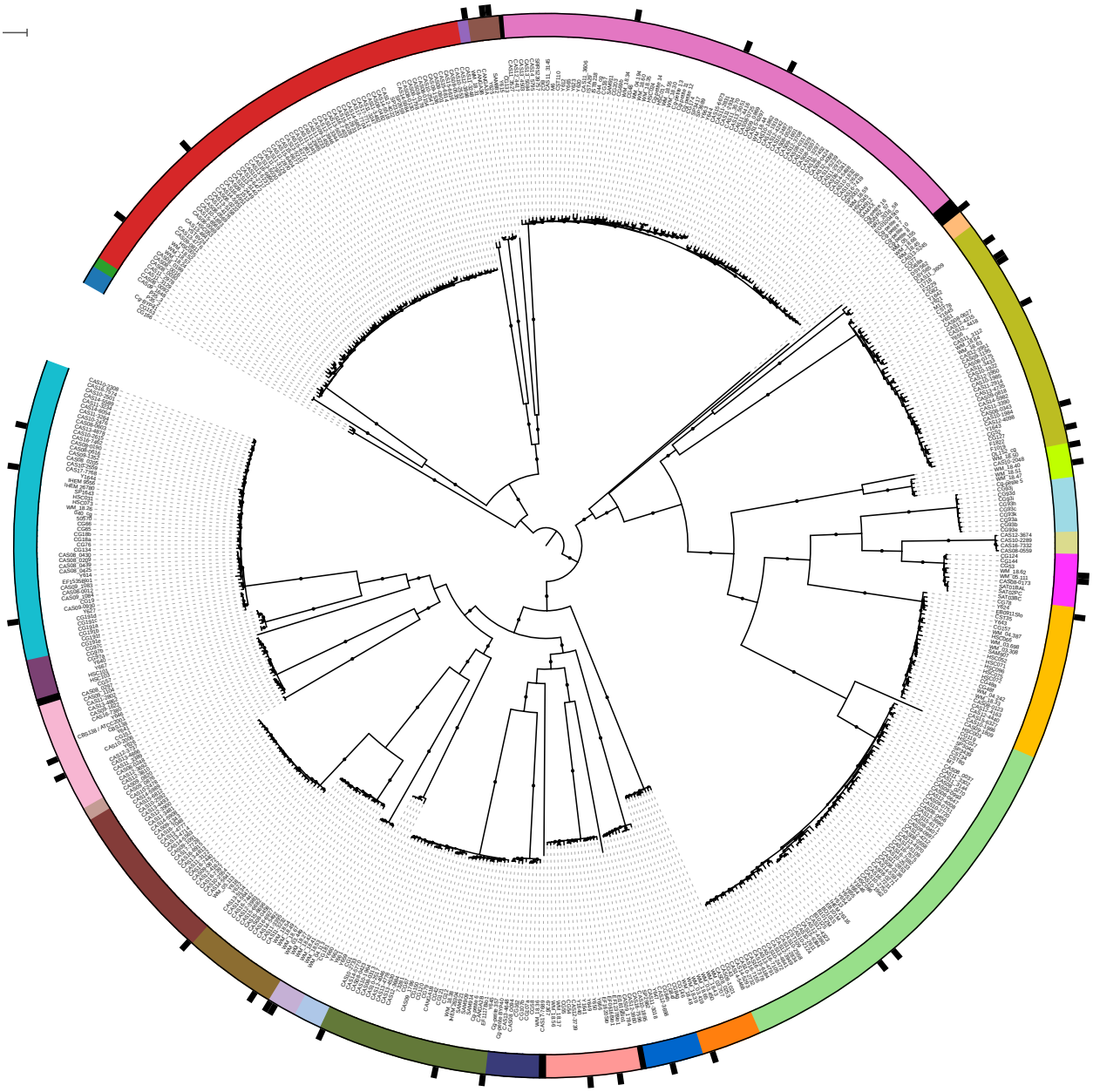


Figure S3.2: Distribution of isolates with karyotypic variation across the *N. glabratus* phylogeny. The colored strip indicates the clusters to which the isolates belong. The outer black bars highlight isolates exhibiting aneuploidies and/or copy number variations.

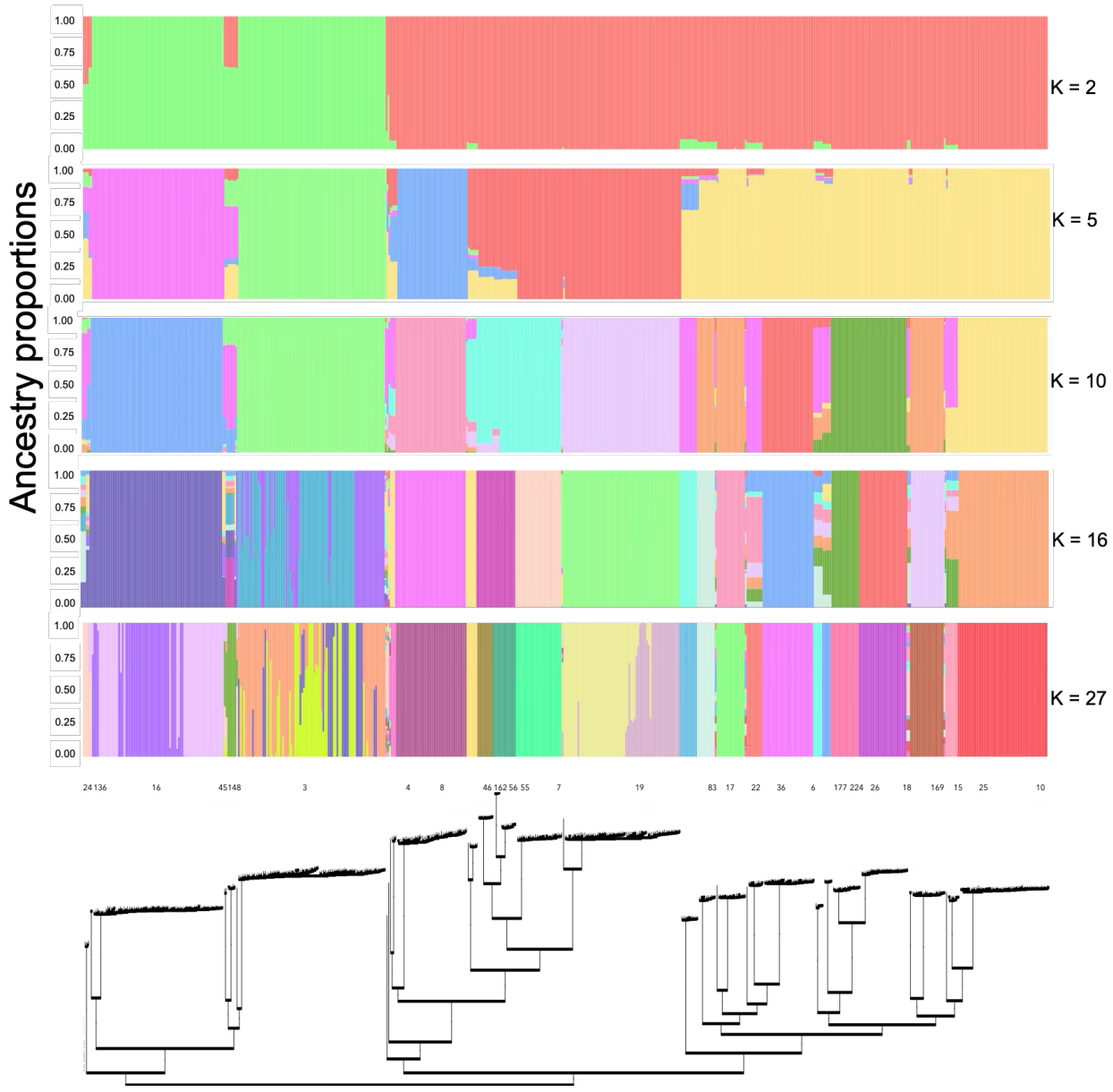


Figure S3.3: Admixture plots for all K-values (K = 2, 5, 10, 16, and 27) with the corresponding phylogeny showing the positions of the isolates.

6.3 Supplementary figures for Chapter 4

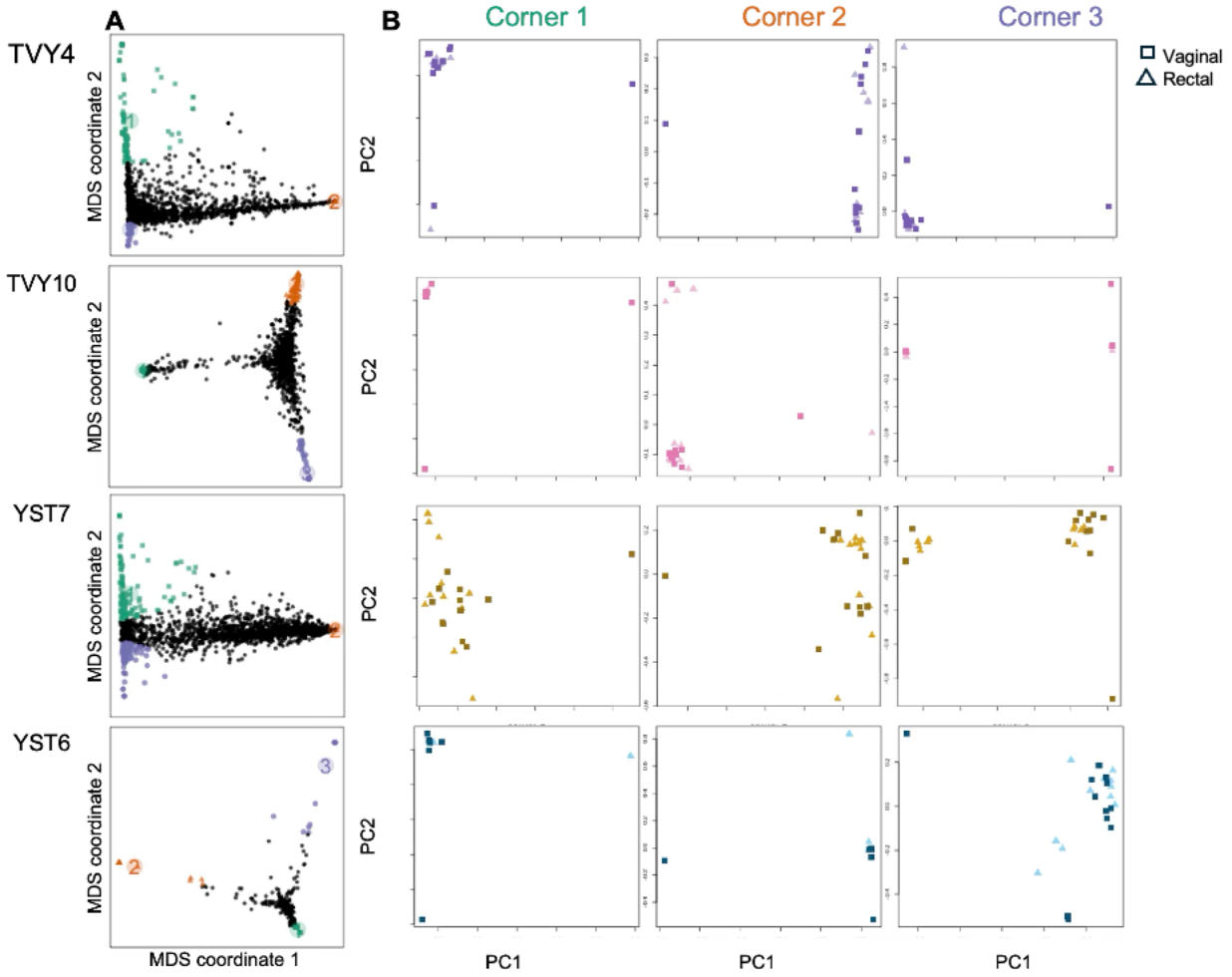
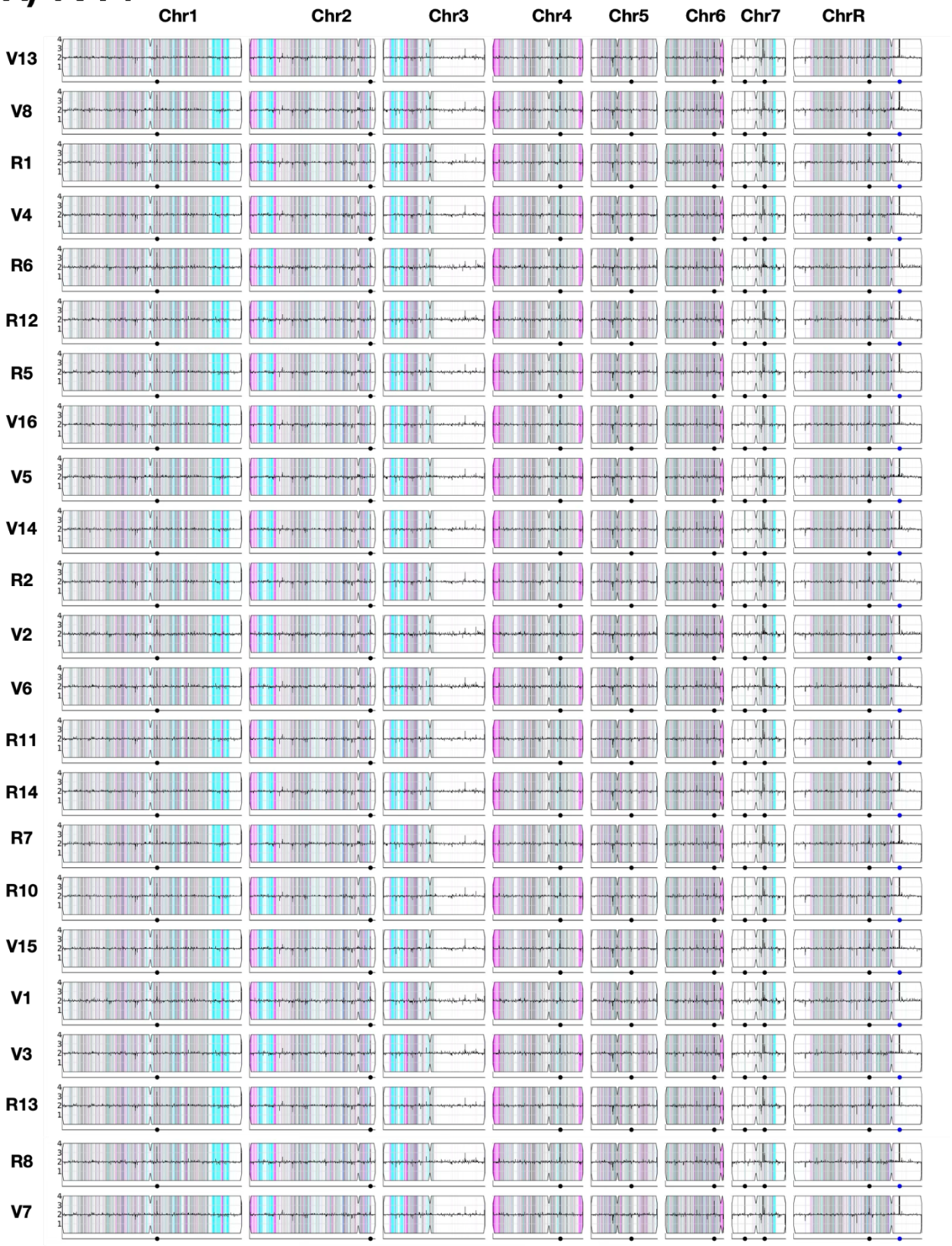
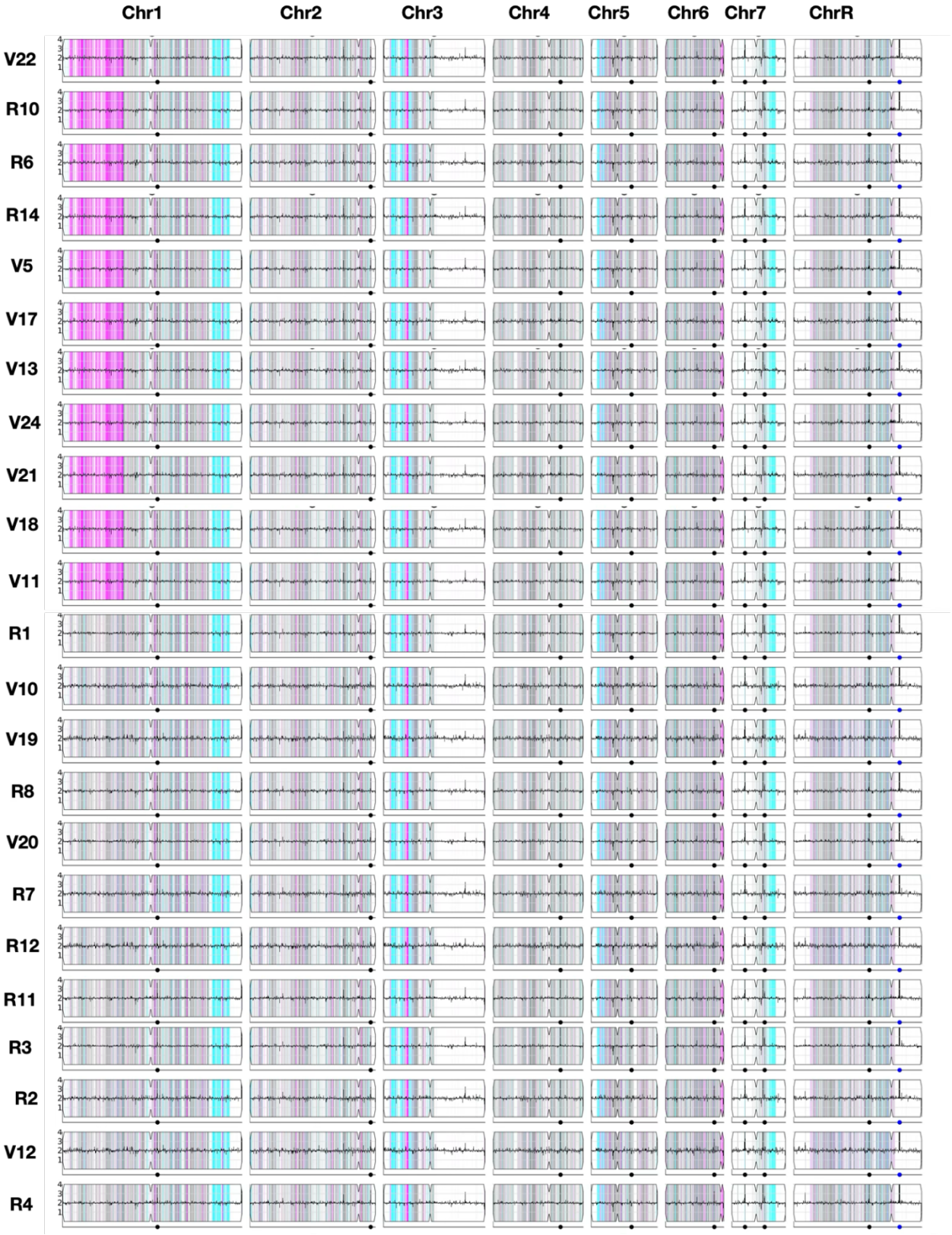


Figure S4.1: Local PCA constructed with the regions of heterogeneity in 5kb windows across the genome. For each participant, three groups of regions of high heterogeneity generated via multidimensional scaling (A) were used in generating the PCA plot (B). This further highlights the overlapping relationship observed in the phylogeny. The pipeline utilizes the multi sample vcf file to divide the genome into discrete windows, allowing for localized analysis of genetic variation. Within each window, patterns of relatedness among individuals are summarized to capture the underlying population structure. These summaries are then compared across all pairs of windows to measure dissimilarity in relatedness, generating a matrix that reflects how patterns vary along the genome. To interpret this high-dimensional matrix, multidimensional scaling (MDS) is applied, providing a visual representation of genomic regions with similar or divergent structure. Finally, windows with similar patterns are grouped together, enabling more accurate visualization of local population structure through principal component analysis.

A) TVY4


Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
ChrR
V13
V8
R1
V4
R6
R12
R5
V16
V5
V14
R2
V2
V6
R11
R14
R7
R10
V15
V1
V3
R13
R8
V7

B) TVY10


Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
ChrR
V22
R10
R6
R14
V5
V17
V13
V24
V21
V18
V11
R1
V10
V19
R8
V20
R7
R12
R11
R3
R2
V12
R4

C) YST7

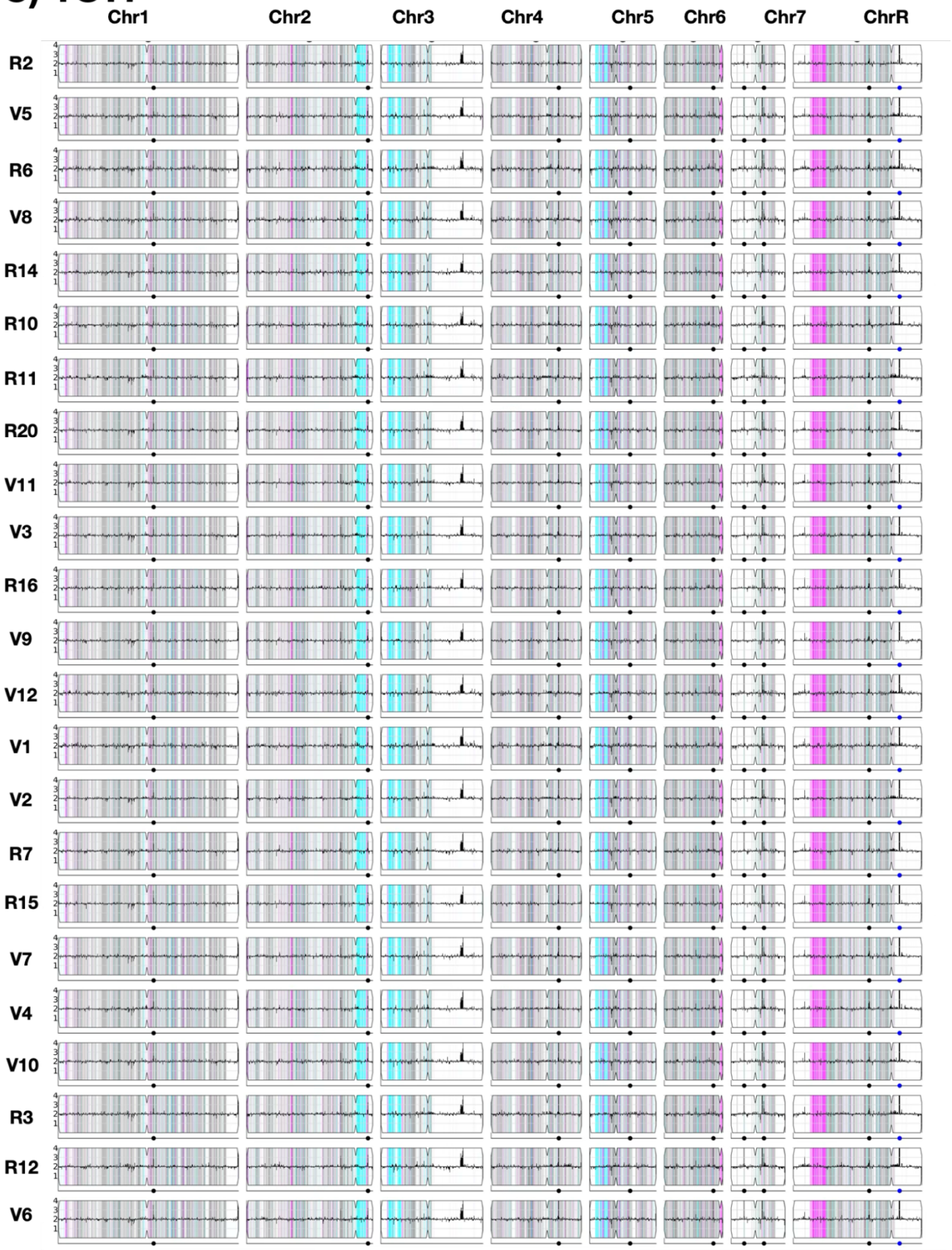

Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
ChrR
R2
V5
R6
V8
R14
R10
R11
R20
V11
V3
R16
V9
V12
V1
V2
R7
R15
V7
V4
V10
R3
R12
V6

Figure S4.2: Copy number variation and loss of heterozygosity were quantified in comparison to the SC5314. A21 reference genome using YMAP (Abbey et al., 2014) from isolates from (A) TVY4, (B) TVY10, and (C) YST7. The order of isolates matches that of the phylogenies shown in Figure 4.2A. The density of single nucleotide polymorphisms is displayed as vertical lines along the length of each chromosome. Color indicates the frequency of SNPs in 5 kb bins relative to the ancestor: white is homozygous in both the ancestor and the isolate of interest, heterozygous SNPs are shown in grey, and homozygous SNPs are color-coded to indicate the retained homolog, with cyan for "AA" and magenta for "BB". Copy number in each bin is depicted by the black line, with the y-axis representing relative copy number; a region with increased reads is visualized as an upward spike in the region. Centromere locations are represented as indentations on the top and bottom of each chromosome box, and the dots along the bottom line mark the positions of major repeat sequences.

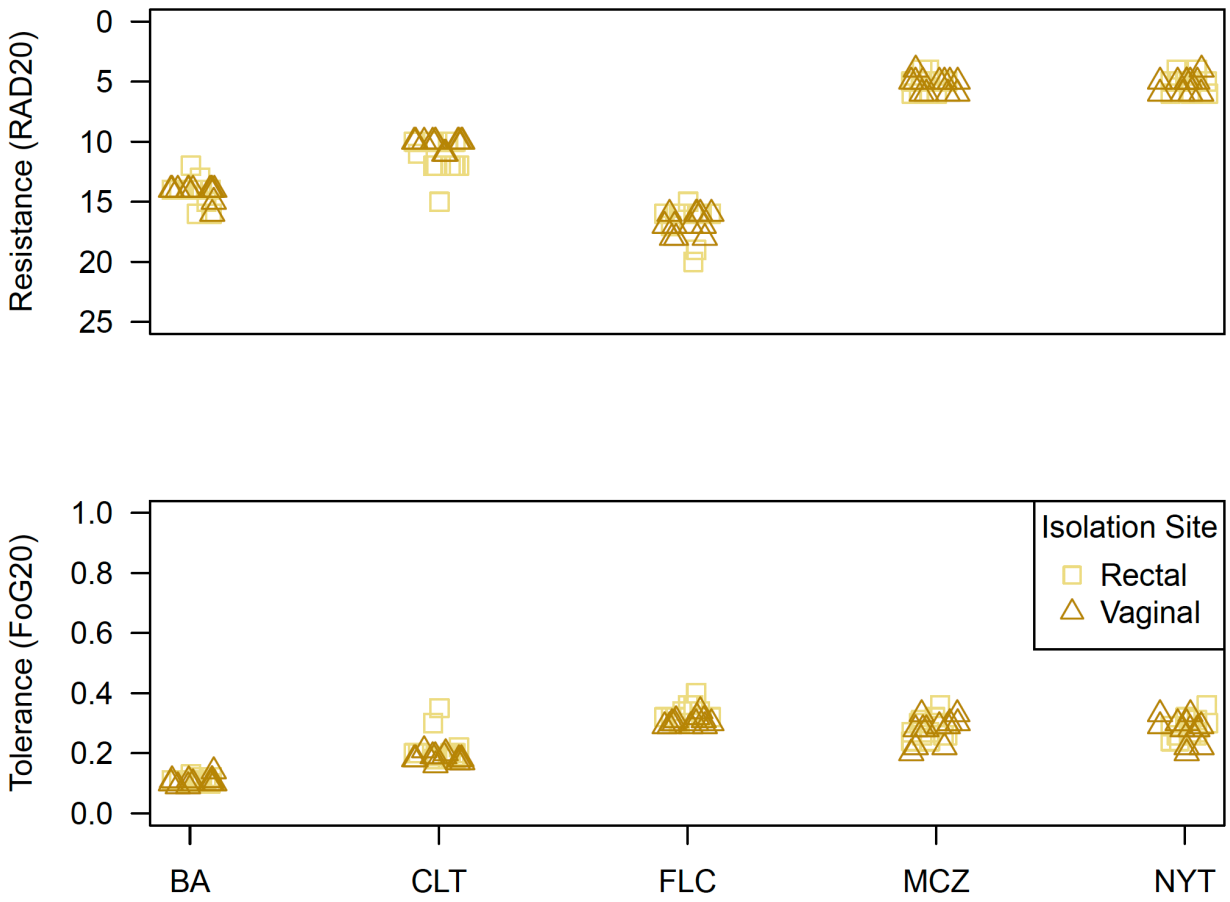


Figure S4.3 Drug response (top: resistance, bottom: tolerance) of 12 vaginal and 12 rectal isolates from YST7

C. albicans isolates to five different drugs (BA: boric acid, CLT: clotrimazole, FLC: fluconazole, MCZ: miconazole, NYT: nystatin,).

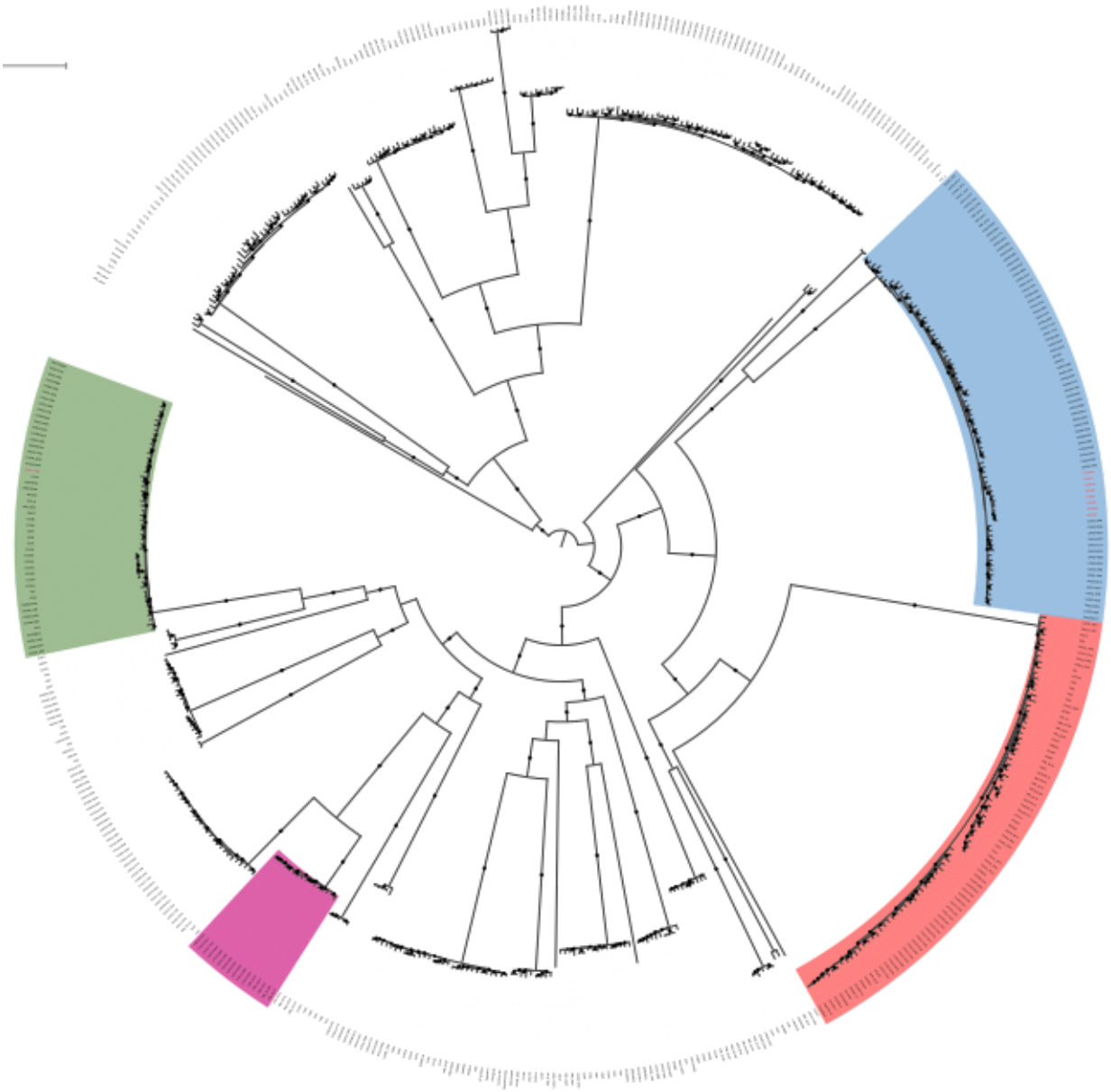


Figure S4.4: Phylogenetic relationship among *N. glabratus* isolates included in genetic variation comparison. The isolates are distributed in 4 clusters colored in the figure**.**

References

Abbes S, Sellami H, Sellami A, et al. (2012) *Candida glabrata* strain relatedness by new microsatellite markers. *European Journal of Clinical Microbiology & Infectious Diseases* 31(1): 83–91.

Abbey DA, Funt J, Lurie-Weinberger MN, et al. (2014) YMAP: a pipeline for visualization of copy number variation and loss of heterozygosity in eukaryotic pathogens. *Genome Medicine* 6(11): 100.

Adamu Bukari A-R, Kukurudz-Gorowski RJ, de Graaf A, et al. (2025) Migration and standing variation in vaginal and rectal yeast populations in recurrent vulvovaginal candidiasis. *mSystems* 10(9): e0015725.

Ajay SS, Parker SCJ, Abaan HO, et al. (2011) Accurate and comprehensive sequencing of personal genomes. *Genome Research* 21(9): 1498–1505.

Alexander DH, Novembre J and Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19(9): 1655–1664.

Alkhars N, Al Jallad N, Wu TT, et al. (2024) Multilocus sequence typing of *Candida albicans* oral isolates reveals high genetic relatedness of mother-child dyads in early life. *PloS One* 19(1): e0290938.

Al-Yasiri MH, Normand A-C, L'Ollivier C, et al. (2016) Opportunistic fungal pathogen *Candida glabrata* circulates between humans and yellow-legged gulls. *Scientific Reports* 6(1): 36157.

Andermann T, Fernandes AM, Olsson U, et al. (2019) Allele phasing greatly improves the phylogenetic utility of ultraconserved elements. *Systematic Biology* 68(1): 32–46.

Anderson FM, Visser ND, Amses KR, 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.

Anderson J, Srikantha T, Morrow B, et al. (1993) Characterization and partial nucleotide sequence of the DNA fingerprinting probe *Ca3* of *Candida albicans*. *Journal of Clinical Microbiology* 31(6): 1472–1480.

Anderson MZ, Saha A, Haseeb A, et al. (2017) A chromosome 4 trisomy contributes to increased fluconazole resistance in a clinical isolate of *Candida albicans*. *Microbiology* 163(6): 856–865.

Aramayo R and Selker EU (2013) *Neurospora crassa*, a model system for epigenetics research. *Cold Spring Harbor Perspectives in Biology* 5(10): a017921.

Arastehfar A, Yazdanpanah S and Bakhtiari M (2021) Epidemiology of candidemia in Shiraz, southern Iran: A prospective multicenter study (2016–2018). *Medical Mycology*, 59(5): 422-430.

Araújo Paulo de Medeiros M, Vieira de Melo AP, Gonçalves SS, et al. (2014) Genetic relatedness among vaginal and anal isolates of *Candida albicans* from women with vulvovaginal candidiasis in north-east Brazil. *Journal of Medical Microbiology* 63(Pt 11): 1436–1445.

Araújo Paulo de Medeiros M, Vieira de Melo AP, Maia de Sousa AM, et al. (2017) Characterization of virulence factors of vaginal and anal isolates of *Candida albicans* sequentially obtained from patients with vulvovaginal candidiasis in north-east Brazil. *Journal de Mycologie Medicale* 27(4): 567–572.

Ashton PM, Thanh LT, Trieu PH, et al. (2019) Three phylogenetic groups have driven the recent population expansion of *Cryptococcus neoformans*. *Nature Communications,* 10(1):2035.

Astvad KMT, Johansen HK, Røder BL, et al. (2018) Update from a 12-Year Nationwide Fungemia Surveillance: Increasing Intrinsic and Acquired Resistance Causes Concern. *Journal of Clinical Microbiology,* 56(4): 10-1128.

Babula O, Lazdane G, Kroica J, et al. (2003) Relation between recurrent vulvovaginal candidiasis, vaginal concentrations of mannose-binding lectin, and a mannose-binding lectin gene polymorphism in Latvian women. *Clinical Infectious Diseases* 37(5): 733-737.

Badrane H, Cheng S, Dupont CL, et al. (2023) Genotypic diversity and unrecognized antifungal resistance among populations of *Candida glabrata* from positive blood cultures. *Nature Communications* 14(1): 5918.

Balaban M, Moshiri N, Mai U, et al. (2019) TreeCluster: Clustering biological sequences using phylogenetic trees. *PloS One* 14(8): e0221068.

Barber AE, Weber M, Kaerger K, et al. (2019) Comparative genomics of serial *Candida glabrata* isolates and the rapid acquisition of echinocandin resistance during therapy. *Antimicrobial agents and chemotherapy* 63(2). *Antimicrobial Agents and Chemotherapy* 63(2): 10-1128.

Barbosa A, Araújo D, Ribeiro E, et al. (2020) *Candida albicans* adaptation on simulated human body fluids under different pH. *Microorganisms* 8(4): 511.

Barchiesi F, Orsetti E, Mazzanti S, et al. (2017) Candidemia in the elderly: What does it change? *PloS One* 12(5): e0176576.

Benedict K, Singleton AL, Jackson BR, et al. (2022) Survey of incidence, lifetime prevalence, and treatment of self-reported vulvovaginal candidiasis, United States, 2020. *BMC Women's Health* 22(1): 147.

Bennett JE, Izumikawa K and Marr KA (2004) Mechanism of increased fluconazole resistance in *Candida glabrata* during prophylaxis. *Antimicrobial agents and chemotherapy* 48(5): 1773–1777.

Bennett RJ (2015) The parasexual lifestyle of *Candida albicans*. *Current opinion in microbiology* 28: 10–17.

Bensasson D, Dicks J, Ludwig JM, et al. (2019) Diverse Lineages of *Candida albicans* Live on Old Oaks. *Genetics* 211(1): 277–288.

Benyas D and Sobel JD (2022) Mixed vaginitis due to bacterial vaginosis and candidiasis. *Journal of lower genital tract disease*. Ovid Technologies (Wolters Kluwer Health): 26(1):68–70.

Berkhout CM (1923) De schimmelgeslachten Monilia, Oidium, Oospora, en Torula [Dissertation]. *University of Utrecht*.

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

Bhattacharjee P (2016) Epidemiology and antifungal susceptibility of *Candida* species in a tertiary care hospital, Kolkata, India. *Current medical mycology* 2(2): 20–27.

Biswas C, Marcelino VR, Van Hal S, et al. (2018) Whole Genome Sequencing of Australian *Candida glabrata* Isolates Reveals Genetic Diversity and Novel Sequence Types. *Frontiers in microbiology* 9: 2946.

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

Blackwell M (2017) Yeasts in insects and other invertebrates. In: *Yeasts in Natural Ecosystems: Diversity* 397–433.

Boisnard S, Zhou Li Y, Arnaise S, et al. (2015) Efficient Mating-Type Switching in *Candida glabrata* Induces Cell Death. *PloS One* 10(10): e0140990.

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

Bolotin-Fukuhara M and Fairhead C (2016) Editorial: *Candida glabrata*, the other yeast pathogen. *FEMS Yeast Research* 16(2): fov116.

Bongomin F, Gago S, Oladele RO, et al. (2017) Global and Multi-National Prevalence of Fungal Diseases—Estimate Precision. *Journal of Fungi* 3(4): 57.

Bordallo-Cardona MÁ, Agnelli C, Gómez-Nuñez A, et al. (2019) MSH2 gene point mutations are not antifungal resistance markers in *Candida glabrata*. *Antimicrobial Agents and Chemotherapy* 63(1): 10-1128.

Borman AM and Johnson EM (2021) Name Changes for Fungi of Medical Importance, 2018 to 2019. *Journal of Clinical Microbiology* 59(2): 10-1128.

Borman AM and Johnson EM (2023) Changes in fungal taxonomy: mycological rationale and clinical implications. *Clinical Microbiology Reviews* 36(4): e0009922.

Borman AM, Petch R, Linton CJ, et al. (2008) *Candida nivariensis*, an emerging pathogenic fungus with multidrug resistance to antifungal agents. *Journal of Clinical Microbiology* 46(3): 933–938.

Bougnoux M-E, Morand S and d'Enfert C (2002) Usefulness of Multilocus Sequence Typing for Characterization of Clinical Isolates of *Candida albicans*. *Journal of Clinical Microbiology* 40(4): 1290.

Bougnoux M-E, Tavanti A, Bouchier C, et al. (2003) Collaborative consensus for optimized multilocus sequence typing of *Candida albicans*. *Journal of Clinical Microbiology* 41(11): 5265–5266.

Bradford LL, Chibucos MC, Ma B, et al. (2017) Vaginal *Candida* spp. genomes from women with vulvovaginal candidiasis. *Pathogens and Disease* 75(6): ftx061.

Braun BR, van Het Hoog M, d'Enfert C, et al. (2005) A human-curated annotation of the *Candida albicans* genome. *PLoS Genetics* 1(1): 36–57.

Bravo GA, Antonelli A, Bacon CD, et al. (2019) Embracing heterogeneity: coalescing the Tree of Life and the future of phylogenomics. *PeerJ.* 7: e6399.

Brawner DL (1991) Comparison between methods for serotyping of *Candida albicans* produces discrepancies in results. *Journal of Clinical Microbiology* 29(5): 1020–1025.

Brawner DL and Cutler JE (1989) Oral *Candida albicans* isolates from nonhospitalized normal carriers, immunocompetent hospitalized patients, and immunocompromised patients with or without acquired immunodeficiency syndrome. *Journal of Clinical Microbiology* 27(6): 1335-1341

Brisse S, Pannier C, Angoulvant A, et al. (2009) Uneven distribution of mating types among genotypes of *Candida glabrata* isolates from clinical samples. *Eukaryotic Cell* 8(3): 287–295.

Brockert PJ, Lachke SA, Srikantha T, et al. (2003) Phenotypic switching and mating type switching of *Candida glabrata* at sites of colonization. *Infection and Immunity* 71(12): 7109–7118.

Brown GD, Denning DW, Gow NAR, et al. (2012) Hidden killers: human fungal infections. *Science Translational Medicine* 4(165): 165rv13-165rv13.

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

Buchalo AS, Nevo E, Wasser SP, et al. (1998) Fungal life in the extremely hypersaline water of the Dead Sea: first records. *Proceedings. Biological sciences* 265(1404). The Royal Society: 1461–1465.

Buesching WJ, Kurek K and Roberts GD (1979) Evaluation of the modified API 20C system for identification of clinically important yeasts. *Journal of clinical microbiology* 9(5). American Society for Microbiology: 565–569.

Bukari A-RA, Kukurudz R, de Graaf A, et al. (2023) Genotypic and phenotypic homogeneity of vaginal and rectal yeast isolates from recurrent vulvovaginal candidiasis. *bioRxiv*. Available at: https://www.biorxiv.org/content/10.1101/2023.07.19.549743v1.full (accessed 27 July 2023).

Burgain A, Tebbji F, Khemiri I, et al. (2020) Metabolic reprogramming in the opportunistic yeast *Candida albicans* in response to hypoxia. *mSphere* 5(1). American Society for Microbiology: 10-1128

Butler G (2014) Comparative genomics of *Candida* species. In: *Candida and* Candidiasis: 27–43.

Caramalac DA, da Silva Ruiz L, de Batista GCM, et al. (2007) *Candida* isolated from vaginal mucosa of mothers and oral mucosa of neonates: occurrence and biotypes concordance. *The Pediatric Infectious Disease Journal* 26(7): 553–557.

Carreté L, Ksiezopolska E, Pegueroles C, et al. (2018) Patterns of genomic variation in the opportunistic pathogen *Candida glabrata* suggest the existence of mating and a secondary association with humans. *Current Biology: CB* 28(1): 15-27.e7.

Carreté L, Ksiezopolska E, Gómez-Molero E, et al. (2019) Genome comparisons of *Candida glabrata* serial clinical isolates reveal patterns of genetic variation in infecting clonal populations. *Frontiers in Microbiology* 10: 112.

Casadevall A and Pirofski L-A (2007) Accidental virulence, cryptic pathogenesis, martians, lost hosts, and the pathogenicity of environmental microbes. *Eukaryotic Cell* 6(12): 2169–2174.

Casadevall A, Steenbergen JN and Nosanchuk JD (2003) "Ready made" virulence and "dual use" virulence factors in pathogenic environmental fungi--the *Cryptococcus neoformans* paradigm. *Current Opinion in Microbiology* 6(4): 332–337.

Cavalheiro M and Teixeira MC (2018) *Candida* Biofilms: Threats, Challenges, and Promising Strategies. *Frontiers of Medicine* 5: 28.

Cavalieri D, Di Paola M, Rizzetto L, et al. (2017) Genomic and Phenotypic Variation in Morphogenetic Networks of Two *Candida albicans* Isolates Subtends Their Different Pathogenic Potential. *Frontiers in Immunology* 8: 1997.

Chapman B, Slavin M, Marriott D, et al. (2017) Changing epidemiology of candidaemia in Australia. *The Journal of antimicrobial chemotherapy* 72(4): 1103–1108.

Chen T, Wagner AS and Reynolds TB (2022) When is it appropriate to take off the mask? Signaling pathways that regulate ß(1,3)-glucan exposure in *Candida albicans*. *Frontiers in Fungal Biology* 3: 842501.

Chenal-Francisque V, Lopez J, Cantinelli T, et al. (2011) Worldwide Distribution of Major Clones ofListeria monocytogenes. *Emerging Infectious Diseases* 17(6): 1110–1112.

Chew KL, Octavia S, Jureen R, et al. (2021) Targeted amplification and MinION nanopore sequencing of key azole and echinocandin resistance determinants of clinically relevant *Candida* spp. from blood culture bottles. *Letters in Applied Microbiology* 73(3): 286–293.

Chew KL, Achik R, Osman NH, et al. (2023) Genomic epidemiology of human candidaemia isolates in a tertiary hospital. *Microbial Genomics* 9(7): 001047

Chibana H, Beckerman JL and Magee PT (2000) Fine-resolution physical mapping of genomic diversity in *Candida albicans*. *Genome Research* 10(12): 1865–1877.

Chow NA, Muñoz JF, Gade L, et al. (2020) Tracing the Evolutionary History and Global Expansion of *Candida auris* Using Population Genomic Analyses. *mBio* 11(2): 10-1128.

Cleary JG, Braithwaite R, Gaastra K, et al. (2015) Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. *bioRxiv*. Available at: https://www.biorxiv.org/content/10.1101/023754 (accessed 1 June 2023).

Coleine C, Stajich JE and Selbmann L (2022) Fungi are key players in extreme ecosystems. *Trends in Ecology & Evolution* 37(6): 517–528.

Cormack BP, Ghori N and Falkow S (1999) An adhesin of the yeast pathogen *Candida glabrata* mediating adherence to human epithelial cells. *Science* 285(5427): 578–582.

Correia A, Sampaio P, James S, et al. (2006) *Candida bracarensis* sp. nov., a novel anamorphic yeast species phenotypically similar to *Candida glabrata*. *International Journal of Systematic and Evolutionary Microbiology* 56(Pt 1): 313–317.

Coste A, Turner V, Ischer F, et al. (2006) A mutation in Tac1p, a transcription factor regulating CDR1 and CDR2, is coupled with loss of heterozygosity at chromosome 5 to mediate antifungal resistance in *Candida albicans*. *Genetics* 172(4): 2139–2156.

Cravener MV, Do E, May G, et al. (2023) Reinforcement amid genetic diversity in the *Candida albicans* biofilm regulatory network. *PLoS Pathogens* 19(1): e1011109.

Cuéllar-Cruz M, Briones-Martin-del-Campo M, Cañas-Villamar I, et al. (2008) High resistance to oxidative stress in the fungal pathogen *Candida glabrata* is mediated by a single catalase, Cta1p, and is controlled by the transcription factors *Yap1p, Skn7p, Msn2p,* and *Msn4p*. *Eukaryotic cell* 7(5). American Society for Microbiology: 814–825.

Cuomo CA, Fanning S, Gujja S, et al. (2019) Genome sequence for *Candida albicans* clinical oral isolate 529L. *Microbiology Resource Announcements* 8(25): 10-1128.

d'Enfert C, Kaune A-K, Alaban L-R, et al. (2021) The impact of the Fungus-Host-Microbiota interplay upon *Candida albicans* infections: current knowledge and new perspectives. *FEMS Microbiology Reviews* 45(3): fuaa060.

da Silva Dantas A, Lee KK, Raziunaite I, et al. (2016) Cell biology of *Candida albicans*-host interactions. *Current opinion in microbiology* 34: 111–118.

Danecek P, Auton A, Abecasis G, et al. (2011) The variant call format and VCFtools. *Bioinformatics* 27(15): 2156–2158.

Danecek P, Bonfield JK, Liddle J, et al. (2021) Twelve years of SAMtools and BCFtools. *GigaScience* 10(2): giab008.

Daniel H-M, Lachance M-A and Kurtzman CP (2014) On the reclassification of species assigned to *Candida*and other anamorphic ascomycetous yeast genera based on phylogenetic circumscription. *Antonie van Leeuwenhoek* 106(1): 67–84.

Davis D (2003) Adaptation to environmental pH in *Candida albicans* and its relation to pathogenesis. *Current Genetics* 44(1): 1–7.

de Melo Pereira GV, Soccol VT, Pandey A, et al. (2014) Isolation, selection and evaluation of yeasts for use in fermentation of coffee beans by the wet process. *International Journal of Food Microbiology* 188: 60–66.

Denning DW (2024) Renaming *Candida glabrata*-A case of taxonomic purity over clinical and public health pragmatism. *PLoS Pathogens* 20(3): e1012055.

Denning DW, Kneale M, Sobel JD, et al. (2018) Global burden of recurrent vulvovaginal candidiasis: a systematic review. *The Lancet Infectious Diseases* 18(11): e339–e347.

Denoncourt A and Downey M (2021) Model systems for studying polyphosphate biology: a focus on microorganisms. *Current Genetics* 67(3): 331–346.

DePristo MA, Banks E, Poplin R, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* 43(5): 491–498.

Dinh N and Bonnefoy N (2024) *Schizosaccharomyces pombe* as a fundamental model for research on mitochondrial gene expression: Progress, achievements and outlooks. *IUBMB life* 76(7): 397–419.

Diogo D, Bouchier C, d'Enfert C, et al. (2009) Loss of heterozygosity in commensal isolates of the asexual diploid yeast *Candida albicans*. *Fungal Genetics and Biology* 46(2): 159–168.

Dodgson AR, Pujol C, Denning DW, et al. (2003) Multilocus sequence typing of *Candida glabrata* reveals geographically enriched clades. *Journal of Clinical Microbiology* 41(12): 5709–5717.

Dodgson AR, Dodgson KJ, Pujol C, et al. (2004) Clade-specific flucytosine resistance is due to a single nucleotide change in the FUR1 gene of *Candida albicans*. *Antimicrobial Agents and Chemotherapy* 48(6): 2223–2227.

Dodgson AR, Pujol C, Pfaller MA, et al. (2005) Evidence for recombination in *Candida glabrata*. *Fungal Genetics and Biology* 42(3): 233–243.

Drell T, Lillsaar T, Tummeleht L, et al. (2013) Characterization of the vaginal micro- and mycobiome in asymptomatic reproductive-age Estonian women. *PloS One* 8(1): e54379.

Dujon B, Sherman D, Fischer G, et al. (2004) Genome evolution in yeasts. *Nature* 430(6995): 35–44.

Dunkel N, Blass J, Rogers PD, et al. (2008) Mutations in the multi-drug resistance regulator MRR1, followed by loss of heterozygosity, are the main cause of MDR1 overexpression in fluconazole-resistant *Candida albicans* strains. *Molecular Microbiology* 69(4): 827–840.

Efros A and Halperin E (2012) Haplotype reconstruction using perfect phylogeny and sequence data. *BMC Bioinformatics* 13 Suppl 6(S6): S3.

Ehrström S, Kornfeld D and Rylander E (2007) Perceived stress in women with recurrent vulvovaginal candidiasis. *Journal of Psychosomatic Obstetrics and Gynaecology* 28(3): 169–176.

El-Din SS, Reynolds MT, Ashbee HR, et al. (2001) An investigation into the pathogenesis of vulvo-vaginal candidosis. *Sexually Transmitted Infections* 77(3): 179–183.

Ene IV, Farrer RA, Hirakawa MP, et al. (2018) Global analysis of mutations driving microevolution of a heterozygous diploid fungal pathogen. *Proceedings of the National Academy of Sciences of the United States of America* 115(37): E8688–E8697.

Ewels P, Magnusson M, Lundin S, et al. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32(19): 3047–3048.

Fabre E, Muller H, Therizols P, et al. (2005) Comparative genomics in hemiascomycete yeasts: evolution of sex, silencing, and subtelomeres. *Molecular Biology and Evolution* 22(4): 856–873.

Farr A, Effendy I, Tirri BF, et al. (2021) Vulvovaginal candidosis (excluding mucocutaneous candidosis): Guideline of the German (DGGG), Austrian (OEGGG) and Swiss (SGGG) society of Gynecology and Obstetrics. *Geburtshilfe und Frauenheilkunde* 81(4): 398–421.

Feil EJ, Cooper JE, Grundmann H, et al. (2003) How clonal is *Staphylococcus aureus*? *Journal of Bacteriology* 185(11): 3307–3316.

Feofilova EP (2001) The Kingdom Fungi: Heterogeneity of Physiological and Biochemical Properties and Relationships with Plants, Animals, and Prokaryotes (Review). *Applied Biochemistry and Microbiology* 37(2): 124–137.

Feri A, Loll-Krippleber R, Commere P-H, et al. (2016) Analysis of Repair Mechanisms following an Induced Double-Strand Break Uncovers Recessive Deleterious Alleles in the *Candida albicans* Diploid Genome. *mBio* 7(5): 10-1128.

Filippidi A, Galanakis E, Maraki S, et al. (2014) The effect of maternal flora on *Candida*colonisation in the neonate. *Mycoses* 57(1): 43–48.

Fischer G and Bradford J (2011) Vulvovaginal candidiasis in postmenopausal women: the role of hormone replacement therapy. *Journal of Lower Genital Tract Disease* 15(4): 263–267.

Fitzpatrick DA, Logue ME, Stajich JE, et al. (2006) A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evolutionary Biology* 6(1): 99.

Flanagan PR, Fletcher J, Boyle H, et al. (2018) Expansion of the TLO gene family enhances the virulence of *Candida*species. *PloS one* 13(7): e0200852.

Fong IW (1994) The rectal carriage of yeast in patients with vaginal candidiasis. *Clinical and investigative medicine.* 17(5): 426–431.

Forche A, Abbey D, Pisithkul T, et al. (2011) Stress alters rates and types of loss of heterozygosity in *Candida albicans*. *mBio* 2(4): 10-1128.

Forche A, Cromie G, Gerstein AC, et al. (2018) Rapid Phenotypic and Genotypic Diversification After Exposure to the Oral Host Niche in *Candida albicans*. *Genetics* 209(3): 725–741.

Forche A, Solis NV, Swidergall M, et al. (2019) Selection of *Candida albicans* trisomy during oropharyngeal infection results in a commensal-like phenotype. *PLoS Genetics* 15(5): e1008137.

Ford CB, Funt JM, Abbey D, et al. (2015) The evolution of drug resistance in clinical isolates of *Candida albicans*. *eLife* 4: e00662.

Fotedar R, Chatting M, Kolecka A, et al. (2022) Communities of culturable yeasts and yeast-like fungi in oligotrophic hypersaline coastal waters of the Arabian Gulf surrounding Qatar. *Antonie van Leeuwenhoek* 115(5): 609–633.

Foxman B, Barlow R, D'Arcy H, et al. (2000) *Candida* vaginitis: self-reported incidence and associated costs. *Sexually Transmitted Diseases* 27(4): 230–235.

Freeman KR, Martin AP, Karki D, et al. (2009) Evidence that chytrids dominate fungal communities in high-elevation soils. *Proceedings of the National Academy of Sciences of the United States of America* 106(43): 18315–18320.

Fricker-Hidalgo H, Vandapel O, Duchesne MA, et al. (1996) Comparison of the new API *Candida* system to the ID 32C system for identification of clinically important yeast species. *Journal of clinical microbiology* 34(7): 1846–1848.

Friedman DZP and Schwartz IS (2019) Emerging Fungal Infections: New Patients, New Patterns, and New Pathogens. *Journal of Fungi* 5(3):67

Fu X-H, Meng F-L, Hu Y, et al. (2008) *Candida albicans*, a distinctive fungal model for cellular aging study. *Aging Cell* 7(5): 746–757.

Fuller Jeff, Dingle TC, Bull A, et al. (2019) Species distribution and antifungal susceptibility of invasive *Candida*isolates from Canadian hospitals: results of the CANWARD 2011–16 study. *The Journal of Antimicrobial Chemotherapy* 74: iv48–iv54.

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

Gabaldón T and Fairhead C (2019) Genomes shed light on the secret life of *Candida glabrata*: not so asexual, not so commensal. *Current Genetics* 65(1): 93–98.

Gabaldón T, Martin T, Marcet-Houben M, et al. (2013) Comparative genomics of emerging pathogens in the *Candida glabrata* clade. *BMC Genomics* 14: 623.

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

Gabaldón T, Gómez-Molero E and Bader O (2020) Molecular Typing of *Candida glabrata*. *Mycopathologia* 185(5): 755–764.

Galagan JE, Henn MR, Ma L-J, et al. (2005) Genomics of the fungal kingdom: insights into eukaryotic biology. *Genome Research* 15(12). Cold Spring Harbor Laboratory: 1620–1631.

Galocha M, Viana R, Pais P, et al. (2022) Genomic evolution towards azole resistance in *Candida glabrata* clinical isolates unveils the importance of CgHxt4/6/7 in azole accumulation. *Communications Biology* 5(1): 1118.

Ge S-H, Xie J, Xu J, et al. (2012) Prevalence of specific and phylogenetically closely related genotypes in the population of *Candida albicans* associated with genital candidiasis in China. *Fungal Genetics and Biology* 49(1): 86–93.

Gerstein AC, Kuzmin A and Otto SP (2014) Loss-of-heterozygosity facilitates passage through Haldane's sieve for *Saccharomyces cerevisiae* undergoing adaptation. *Nature Communications* 5: 3819.

Gerstein AC, Rosenberg A, Hecht I, et al. (2016) diskImageR: quantification of resistance and tolerance to antimicrobial drugs using disk diffusion assays. *Microbiology* 162(7): 1059–1068.

Ghannoum MA, Jurevic RJ, Mukherjee PK, et al. (2010) Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. *PLoS Pathogens* 6(1): e1000713.

Giblin L, Edelmann A, Zhang N, et al. (2001) A DNA polymorphism specific to *Candida albicans* strains exceptionally successful as human pathogens. *Gene* 272(1–2): 157–164.

Giraldo PC, Polpeta NC, Juliato CRT, et al. (2012) Evaluation of sexual function in Brazilian women with recurrent vulvovaginal candidiasis and localized provoked vulvodynia. *The Journal of Sexual Medicine* 9(3): 805–811.

Gnaien M, Maufrais C, Rebai Y, et al. (2024) A gain-of-function mutation in zinc cluster transcription factor Rob1 drives *Candida albicans* adaptive growth in the cystic fibrosis lung environment. *PLoS Pathogens* 20(4): e1012154.

Gonçalves VN, Cantrell CL, Wedge DE, et al. (2016) Fungi associated with rocks of the Atacama Desert: taxonomy, distribution, diversity, ecology and bioprospection for bioactive compounds: Fungi associated with rocks of the Atacama Desert. *Environmental Microbiology* 18(1): 232–245.

Gong J, Chen X-F, Fan X, et al. (2023) Emergence of antifungal resistant subclades in the global predominant phylogenetic population of *Candida albicans*. *Microbiology Spectrum* 11(1): e0380722.

Guinea J, Mezquita S, Gómez A, et al. (2021) Whole genome sequencing confirms *Candida albicans* and *Candida parapsilosis* microsatellite sporadic and persistent clones causing outbreaks of candidemia in neonates. *Medical Mycology* 60(1): myab068.

Guo X, Zhang R, Li Y, et al. (2020) Understand the genomic diversity and evolution of fungal pathogen *Candida glabrata* by genome-wide analysis of genetic variations. *Methods* 176: 82–90.

Gupta A, Jordan IK and Rishishwar L (2017) stringMLST: a fast k-mer based tool for multilocus sequence typing. *Bioinformatics* 33(1): 119–121.

Guzel AB, Ilkit M, Akar T, et al. (2011) Evaluation of risk factors in patients with vulvovaginal candidiasis and the value of chromID *Candida* agar versus CHROMagar *Candida* for recovery and presumptive identification of vaginal yeast species. *Medical Mycology* 49(1): 16–25.

Guzmán B, Lachance M-A and Herrera CM (2013) Phylogenetic analysis of the angiosperm-floricolous insect-yeast association: have yeast and angiosperm lineages co-diversified? *Molecular Phylogenetics and Evolution* 68(2): 161–175.

Haase JK, Didelot X, Lecuit M, et al. (2014) The ubiquitous nature of Listeria monocytogenes clones: a large-scale Multilocus Sequence Typing study: MLST of *L*. *monocytogenes*. *Environmental Microbiology* 16(2): 405–416.

Haller BC and Messer PW (2023) SLiM 4: Multispecies Eco-evolutionary modeling. *The American Naturalist* 201(5): E127–E139.

Hameed S, Hans S, Singh S, et al. (2021) Revisiting the vital drivers and mechanisms of β-glucan masking in human fungal pathogen, *Candida albicans*. *Pathogens* 10(8): 942.

Hamlin JAP, Dias GB, Bergman CM, et al. (2019) Phased Diploid Genome Assemblies for Three Strains of *Candida albicans* from Oak Trees. *G3: Genes, Genomes, Genetic*s, 9(11): 3547-3554.

Hasenclever HF and Mitchell WO (1961a) Antigenic studies of Candida. I. Observation of two antigenic groups in *Candida albicans*. *Journal of Bacteriology* 82: 570–573.

Hasenclever HF and Mitchell WO (1961b) Antigenic studies of *Candida*. III. Comparative pathogenicity of *Candida albicans* group A, group B, and *Candida stellatoidea*. *Journal of Bacteriology* 82: 578–581.

Håvelsrud OE and Gaustad P (2017) Draft Genome Sequences of *Candida glabrata* Isolates 1A, 1B, 2A, 2B, 3A, and 3B. *Genome Announcements* 5(10): e00328-16.

Hawksworth DL (2001) The magnitude of fungal diversity: the 1.5 million species estimate revisited. *Mycological research* 105(12): 1422–1432.

Hawksworth DL and Lücking R (2017) Fungal Diversity Revisited: 2.2 to 3.8 Million Species. *Microbiology Spectrum* 5(4): 10-1128.

He X, Kusuya Y, Hagiwara D, et al. (2024) Genomic diversity of the pathogenic fungus Aspergillus fumigatus in Japan reveals the complex genomic basis of azole resistance. *Communications biology* 7(1): 274.

Healey KR, Zhao Y, Perez WB, et al. (2016) Prevalent mutator genotype identified in fungal pathogen *Candida glabrata* promotes multi-drug resistance. *Nature Communications* 7(1): 11128.

Helmstetter N, Chybowska AD, Delaney C, et al. (2022) Population genetics and microevolution of clinical *Candida glabrata* reveals recombinant sequence types and hyper-variation within mitochondrial genomes, virulence genes, and drug targets. *Genetics* 221(1): iyac031.

Hirakawa MP, Martinez DA, Sakthikumar S, et al. (2015) Genetic and phenotypic intra-species variation in *Candida albicans*. *Genome Research* 25(3): 413–425.

Hoffman CS, Wood V and Fantes PA (2015) An ancient yeast for young geneticists: A Primer on the *Schizosaccharomyces pombe* model system. *Genetics* 201(2): 403–423.

Honda S, Eusebio-Cope A, Miyashita S, et al. (2020) Establishment of *Neurospora crassa* as a model organism for fungal virology. *Nature Communications* 11(1): 5627.

Hoyer LL and Cota E (2016) *Candida albicans* agglutinin-like sequence (Als) family vignettes: A review of Als protein structure and function. *Frontiers in Microbiology* 7: 280.

Hunter PR and Fraser C (1987) Use of modified resistogram to type *Candida albicans* isolated from cases of vaginitis and from faeces in the same geographical area. *Journal of Clinical Pathology* 40(10). BMJ: 1159–1161.

Hunter PR, Fraser CA and Mackenzie DW (1989) Morphotype markers of virulence in human candidal infections. *Journal of Medical Microbiology* 28(2): 85–91.

Iqbal Z, Caccamo M, Turner I, et al. (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics* 44(2): 226–232.

Iracane E, Arias-Sardá C, Maufrais C, et al. (2024) Identification of an active RNAi pathway in *Candida albicans*. *Proceedings of the National Academy of Sciences* 121(17): e2315926121.

Iraqui I, Garcia-Sanchez S, Aubert S, et al. (2005) The Yak1p kinase controls expression of adhesins and biofilm formation in *Candida glabrata* in a Sir4p-dependent pathway: Yak1p of *Candida glabrata*. *Molecular microbiology* 55(4): 1259–1271.

Jampol LM, Sung J, Walker JD, et al. (1996) Choroidal neovascularization secondary to *Candida albicans* chorioretinitis. *American journal of ophthalmology* 121(6): 643–649.

Jay A, Jordan DF, Gerstein A, et al. (2025) The role of gene copy number variation in antimicrobial resistance in human fungal pathogens. *Antimicrobials and Resistance* 3(1): 1.

Jiang TT, Shao T-Y, Ang WXG, et al. (2017) Commensal fungi recapitulate the protective benefits of intestinal bacteria. *Cell Host & Microbe* 22(6): 809-816.e4.

Jolley KA, Bray JE and Maiden MCJ (2018) Open-access bacterial population genomics: BIGSdb software, the Pubmlst.org website and their applications. *Wellcome Open Research* 3: 124.

Jones T, Federspiel NA, Chibana H, et al. (2004) The diploid genome sequence of *Candida albicans*. *Proceedings of the National Academy of Science* 101(19): 7329–7334.

Kabir MA, Hussain MA and Ahmad Z (2012) *Candida albicans*: A model organism for studying fungal pathogens. *ISRN Microbiology* 2012(1): 538694.

Kakade P, Sircaik S, Maufrais C, et al. (2023) Aneuploidy and gene dosage regulate filamentation and host colonization by *Candida albicans*. *Proceedings of the National Academy of Sciences* 120(11): e2218163120.

Kaplan E, Aktaş D, Önder Ş, et al. (2019) Mating genotypes and susceptibility profiles of clinical isolates of *Candida glabrata* from Turkey. *Mycoses* 62(9): 796–802.

Karathia H, Vilaprinyo E, Sorribas A, et al. (2011) *Saccharomyces cerevisiae* as a model organism: a comparative study. *PloS One* 6(2): e16015.

Kashem SW and Kaplan DH (2016) Skin Immunity to *Candida albicans*. *Trends in Immunology* 37(7): 440–450.

Kashem SW, Igyarto BZ, Gerami-Nejad M, et al. (2015) *Candida albicans* morphology and dendritic cell subsets determine T helper cell differentiation. *Immunity* 42(2): 356–366.

Katsipoulaki M, Stappers MHT, Malavia-Jones D, et al. (2024) *Candida albicans* and *Candida glabrata*: global priority pathogens. *Microbiology and Molecular Biology Reviews: MMBR* 88(2): e0002123.

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

Kazmerski TM, Sawicki GS, Miller E, et al. (2018) Sexual and reproductive health behaviors and experiences reported by young women with cystic fibrosis. *Journal of Cystic Fibrosis:* 17(1): 57–63.

Kennedy MA and Sobel JD (2010) Vulvovaginal candidiasis caused by non-albicans *Candida* species: New insights. *Current Infectious Disease Reports* 12(6): 465–470.

Kidd SE, Abdolrasouli A and Hagen F (2023) Fungal nomenclature: managing change is the name of the game. *Open Forum Infectious Diseases* 10(1): ofac559.

Kitaya S, Kanamori H, Katori Y, et al. (2023) Clinical features and outcomes of persistent candidemia caused by *Candida albicans* versus non-albicans *Candida* species: a focus on antifungal resistance and follow-up blood cultures. *Microorganisms* 11(4): 928.

Knoke M and Bernhardt H (2006) The first description of an oesophageal candidosis by Bernhard von Langenbeck in 1839. *Mycoses* 49(4): 283–287.

Koboldt DC (2020) Best practices for variant calling in clinical sequencing. *Genome Medicine* 12(1): 91.

Kolondra A, Labedzka-Dmoch K, Wenda JM, et al. (2015) The transcriptome of *Candida albicans* mitochondria and the evolution of organellar transcription units in yeasts. *BMC Genomics* 16(1): 827.

Kord M, Salehi M and Khodavaisy S (2020) Epidemiology of yeast species causing bloodstream infection in Tehran, Iran (2015–2017); superiority of 21-plex PCR over the Vitek 2 system for yeast identification. *Journal of Medical Microbiology*, 69(5): 712-720.

Korunes KL and Samuk K (2021) pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources* 21(4): 1359–1368.

Koszul R, Malpertuy A, Frangeul L, et al. (2003) The complete mitochondrial genome sequence of the pathogenic yeast *Candida* (*Torulopsis*) *glabrata*. *FEBS Letters* 534(1–3): 39–48.

Krusche P, Trigg L, Boutros PC, et al. (2019) Best practices for benchmarking germline small-variant calls in human genomes. *Nature Biotechnology* 37(5): 555–560.

Ksiezopolska E, Schikora-Tamarit MÀ, Carlos Nunez-Rodriguez J, et al. (2024) Long-term stability of acquired drug resistance and resistance associated mutations in the fungal pathogen *Nakaseomyces glabratus* (*Candida glabrata*). *Frontiers in Cellular and Infection Microbiology* 14: 1416509.

Kukurudz RJ, Chapel M, Wonitowy Q, et al. (2022) Acquisition of cross-azole tolerance and aneuploidy in *Candida albicans* strains evolved to posaconazole. *G3* 12(9):p.jkac156.

Kurtzman C (2003) Phylogenetic circumscription of , and other members of the Saccharomycetaceae, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygotorulaspora*. *FEMS Yeast Research* 4(3): 233–245.

Kurtzman CP (2011) Discussion of teleomorphic and anamorphic ascomycetous yeasts and yeast-like taxa. In: *The Yeasts*: 293–307.

Kurtzman CP and Robnett CJ (1994) Synonymy of the yeast genera *Wingea* and *Debaryomyces*. *Antonie van Leeuwenhoek* 66(4): 337–342.

Kurtzman CP and Robnett CJ (2003) Phylogenetic relationships among yeasts of the "*Saccharomyces* complex" determined from multigene sequence analyses. *FEMS Yeast Research* 3(4): 417–432.

Kurtzman CP and Suzuki M (2010) Phylogenetic analysis of ascomycete yeasts that form coenzyme Q-9 and the proposal of the new genera *Babjeviella*, *Meyerozyma*, *Millerozyma*, *Priceomyces*, and *Scheffersomyces*. *Mycoscience* 51(1): 2–14.

Kuthan R (2025) *Nakaseomyces glabratus* (*Candida glabrata*) MLST Genotypes in Central Poland. *International Journal of Molecular Sciences* 26(9): 4407.

Lachke SA, Joly S, Daniels K, et al. (2002) Phenotypic switching and filamentation in *Candida glabrata*. *Microbiology* 148(Pt 9): 2661–2674.

Lass-Flörl C, Mayr A, Aigner M, et al. (2018) A nationwide passive surveillance on fungal infections shows a low burden of azole resistance in molds and yeasts in Tyrol, Austria. *Infection* 46(5): 701–704.

Latgé J-P and Chamilos G (2019) *Aspergillus fumigatus* and aspergillosis in 2019. *Clinical Microbiology Reviews* 33(1): 10-1128.

Legrand M, Jaitly P, Feri A, et al. (2019) *Candida albicans*: An emerging yeast model to study eukaryotic genome plasticity. *Trends in Genetics*: 292–307.

Lehmann PF, Kemker BJ, Hsiao CB, et al. (1989) Isoenzyme biotypes of *Candida* species. *Journal of Clinical Microbiology* 27(11): 2514–2521.

Lemberg C, Martinez de San Vicente K, Fróis-Martins R, et al. (2022) *Candida albicans* commensalism in the oral mucosa is favoured by limited virulence and metabolic adaptation. *PLoS Pathogens* 18(4): e1010012.

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.

Lew-Smith J, Binkley J and Sherlock G (2025) The *Candida* Genome Database: annotation and visualization updates. *Genetics* 229(3): iyaf001.

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]*. Available at: http://arxiv.org/abs/1303.3997.

Li H and Ralph P (2019) Local PCA Shows How the Effect of Population Structure Differs Along the Genome. *Genetics* 211(1): 289–304.

Li H, Handsaker B, Wysoker A, et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16): 2078–2079.

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

Li XV, Leonardi I, Putzel GG, et al. (2022) Immune regulation by fungal strain diversity in inflammatory bowel disease. *Nature* 603(7902): 672–678.

Liang S-H, Anderson MZ, Hirakawa MP, et al. (2019) Hemizygosity Enables a Mutational Transition Governing Fungal Virulence and Commensalism. *Cell Host & Microbe* 25(3): 418-431.e6.

Lin C-Y, Chen Y-C, Lo H-J, et al. (2007) Assessment of *Candida glabrata* strain relatedness by pulsed-field gel electrophoresis and multilocus sequence typing. *Journal of Clinical Microbiology* 45(8): 2452–2459.

Lipperheide V, Quindós G, Jiménez Y, et al. (1996) *Candida* biotypes in patients with oral leukoplakia and lichen planus. *Candida* biotypes in leukoplakia and lichen planus. *Mycopathologia* 134(2): 75–82.

Liu K, Linder CR and Warnow T (2011) RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PloS One* 6(11): e27731.

Lockhart SR, Fritch JJ, Meier AS, et al. (1995) Colonizing populations of *Candida albicans* are clonal in origin but undergo microevolution through C1 fragment reorganization as demonstrated by DNA fingerprinting and C1 sequencing. *Journal of Clinical Microbiology* 33(6): 1501–1509.

Lockhart SR, Pujol C, Daniels KJ, et al. (2002) In *Candida albicans*, white-opaque switchers are homozygous for mating type. *Genetics* 162(2): 737–745.

Lohse MB and Johnson AD (2009) White-opaque switching in *Candida albicans*. *Current Opinion in Microbiology* 12(6): 650–654.

Lott TJ, Frade JP and Lockhart SR (2010) Multilocus sequence type analysis reveals both clonality and recombination in populations of *Candida glabrata* bloodstream isolates from U.S. surveillance studies. *Eukaryotic Cell* 9(4): 619–625.

Lubin IM, Aziz N, Babb LJ, et al. (2017) Principles and recommendations for standardizing the use of the next-generation sequencing variant file in clinical settings. *The Journal of Molecular Diagnostics* 19(3): 417–426.

Lutzoni F, Kauff F, Cox CJ, et al. (2004) Assembling the fungal tree of life: progress, classification, and evolution of subcellular traits. *American Journal of Botany* 91(10): 1446–1480.

MacCallum DM, Castillo L, Nather K, et al. (2009) Property differences among the four major *Candida albicans* strain clades. *Eukaryotic Cell* 8(3): 373–387.

Maciel NO, Johann S, Brandão LR, et al. (2019) Occurrence, antifungal susceptibility, and virulence factors of opportunistic yeasts isolated from Brazilian beaches. *Memorias do Instituto Oswaldo Cruz* 114: e180566.

Makanjuola O, Bongomin F and Fayemiwo SA (2018) An update on the roles of non-albicans *Candida* species in vulvovaginitis. *Journal of Fungi* 4(4): 121.

Marcet-Houben M, Księżopolska E and Gabaldón T (2024) Chromosome level assemblies of *Nakaseomyces* (*Candida*) *bracarensis* uncover two distinct clades and define its adhesin repertoire. *BMC Genomics* 25(1): 1053.

Marco F, Lockhart SR, Pfaller MA, et al. (1999) Elucidating the origins of nosocomial infections with *Candida albicans* by DNA fingerprinting with the complex probe *Ca3*. *Journal of Clinical Microbiology* 37(9): 2817–2828.

Mårdh P-A, Novikova N and Stukalova E (2003) Colonisation of extragenital sites by *Candida* in women with recurrent vulvovaginal candidosis. *BJOG: an International Journal of Obstetrics and Gynaecology* 110(10): 934–937.

Mardis ER (2012) Applying next-generation sequencing to pancreatic cancer treatment. *Nature Reviews. Gastroenterology & Hepatology* 9(8): 477–486.

Marichal P, Vanden Bossche H, Odds FC, et al. (1997) Molecular biological characterization of an azole-resistant *Candida glabrata* isolate. *Antimicrobial Agents and Chemotherapy* 41(10): 2229–2237.

Martínez-Jiménez V, Ramírez-Zavaleta CY, Orta-Zavalza E, et al. (2013) Sir3 Polymorphisms in *Candida glabrata* clinical isolates. *Mycopathologia* 175(3–4): 207–219.

Marton T, Chauvel M, Feri A, et al. (2021) Factors that influence bidirectional long-tract homozygosis due to double-strand break repair in *Candida albicans*. *Genetics* 218(1): iyab028

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

McCreight MC, Warnock DW and Martin MV (1985) Resistogram typing of *Candida albicans* isolates from oral and cutaneous sites in irradiated patients. *Sabouraudia* 23(6): 403-406.

McTaggart LR, Cabrera A, Cronin K, et al. (2020) Antifungal susceptibility of clinical yeast isolates from a large Canadian reference laboratory and application of whole-genome sequence analysis to elucidate mechanisms of acquired resistance. *Antimicrobial Agents and Chemotherapy* 64(9):10-1128..

Millard SP (2013) *EnvStats: An R Package for Environmental Statistics*. 2nd ed. New York, NY: Springer.

Miller LG, Hajjeh RA and Edwards JE Jr (2001) Estimating the cost of nosocomial candidemia in the united states. *Clinical infectious diseases.* 32(7): 1110-1110*.*

Milne J and Warnock D (1979) Effect of simultaneous oral and vaginal treatment on the rate of cure and relapse in vaginal candidosis. *The British Journal of Venereal Diseases* 55: 362–365.

Mishra A, Solis NV, Dietz SM, et al. (2025) Strain background interacts with chromosome 7 aneuploidy to determine commensal and virulence phenotypes in *Candida albicans*. *PLoS Genetics* 21(6): e1011650.

Mixão V and Gabaldón T (2020) Genomic evidence for a hybrid origin of the yeast opportunistic pathogen *Candida albicans*. *BMC Biology* 18(1): 48.

Mixão V, Saus E, Boekhout T, et al. (2021) Extreme diversification driven by parallel events of massive loss of heterozygosity in the hybrid lineage of *Candida albicans*. *Genetics* 217(2): iyaa004.

Mohammadi S, Leduc A, Charette SJ, et al. (2023) Amino acid substitutions in specific proteins correlate with farnesol unresponsiveness in *Candida albicans*. *BMC genomics* 24(1): 93.

Muller H, Hennequin C, Gallaud J, et al. (2008) The asexual yeast *Candida glabrata* maintains distinct a and alpha haploid mating types. *Eukaryotic Cell* 7(5): 848–858.

Mutreja A, Kim DW, Thomson NR, et al. (2011) Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* 477(7365): 462–465.

Muzzey D, Schwartz K, Weissman JS, et al. (2013) Assembly of a phased diploid *Candida albicans* genome facilitates allele-specific measurements and provides a simple model for repeat and indel structure. *Genome Biology* 14(9): R97.

Nagahama T, Takahashi E, Nagano Y, et al. (2011) Molecular evidence that deep-branching fungi are major fungal components in deep-sea methane cold-seep sediments: Deep branching fungi are dominant in deep-sea. *Environmental Microbiology* 13(8): 2359–2370.

Naranjo-Ortiz MA and Gabaldón T (2019) Fungal evolution: diversity, taxonomy and phylogeny of the Fungi. *Biological reviews of the Cambridge Philosophical Society* 94(6): 2101–2137.

Nash AK, Auchtung TA, Wong MC, et al. (2017) The gut mycobiome of the Human Microbiome Project healthy cohort. *Microbiome* 5(1): 153.

Nielsen J (2019) Yeast systems biology: Model organism and cell factory. *Biotechnology Journal* 14(9): e1800421.

Noble SM, Gianetti BA and Witchley JN (2017) *Candida albicans* cell-type switching and functional plasticity in the mammalian host. *Nature Reviews. Microbiology* 15(2): 96–108.

Nyirjesy P, Peyton C, Weitz MV, et al. (2006) Causes of chronic vaginitis: analysis of a prospective database of affected women. *Obstetrics and Gynecology.* 108(5): 1185-1191.

O'Brien CE, Oliveira-Pacheco J, Ó Cinnéide E, et al. (2021) Population genomics of the pathogenic yeast *Candida tropicalis* identifies hybrid isolates in environmental samples. *PLoS Pathogens* 17(3): e1009138.

O'Brien D, Stevens NT, Lim CH, et al. (2011) *Candida* infection of the central nervous system following neurosurgery: a 12-year review. *Acta Neurochirurgica* 153(6): 1347–1350.

O'Connor MI and Sobel JD (1986) Epidemiology of recurrent vulvovaginal candidiasis: identification and strain differentiation of *Candida albicans*. *The Journal of Infectious Diseases* 154(2): 358–363.

Odds EC (1997) Switch of phenotype as an escape mechanism of the intruder. *Mycoses* 40 (Supp): 9–12.

Odds FC (2009) In *Candida albicans*, resistance to flucytosine and terbinafine is linked to MAT locus homozygosity and multilocus sequence typing clade 1. *FEMS Yeast Research* 9(7): 1091–1101.

Odds FC (2010) Molecular phylogenetics and epidemiology of *Candida albicans*. *Future Microbiology* 5(1): 67–79.

Odds FC and Abbott AB (1980) A simple system for the presumptive identification of *Candida albicans* and differentiation of strains within the species. *Sabouraudia* 18(4): 301–317.

Odds FC and Abbott AB (1983) Modification and extension of tests for differentiation of *Candida*species and strains. *Sabouraudia* 21(1): 79–81.

Odds FC and Jacobsen MD (2008) Multilocus sequence typing of pathogenic *Candida* species. *Eukaryotic Cell* 7(7): 1075–1084.

Odds FC, Kibbler CC, Walker E, et al. (1989) Carriage of *Candida* species and *C. albicans* biotypes in patients undergoing chemotherapy or bone marrow transplantation for haematological disease. *Journal of Clinical Pathology* 42(12): 1259–1266.

Odds FC, Davidson AD, Jacobsen MD, et al. (2006) *Candida albicans* strain maintenance, replacement, and microvariation demonstrated by multilocus sequence typing. *Journal of Clinical Microbiology* 44(10): 3647–3658.

Odds FC, Bougnoux M-E, Shaw DJ, et al. (2007) Molecular phylogenetics of *Candida albicans*. *Eukaryotic Cell* 6(6): 1041–1052.

Oliver JC, Ferreira CBRJ, Silva NC, et al. (2019) *Candida* spp. and phagocytosis: multiple evasion mechanisms. *Antonie van Leeuwenhoek* 112(10): 1409–1423.

Olson ND, Wagner J, Dwarshuis N, et al. (2023) Variant calling and benchmarking in an era of complete human genome sequences. *Nature Reviews. Genetics* 24(7): 464–483.

Opulente DA, Langdon QK, Buh KV, et al. (2019) Pathogenic budding yeasts isolated outside of clinical settings. *FEMS Yeast Research* 19(3): foz032.

Oren A and Gunde-Cimerman N (2012) Fungal life in the dead sea. *Progress in Molecular and Subcellular Biology* 53: 115–132.

Ortiz EM (2019) *Vcf2phylip v2.0: Convert a VCF Matrix into Several Matrix Formats for Phylogenetic Analysis*. Available at: https://zenodo.org/record/2540861.

Oxman DA, Chow JK, Frendl G, et al. (2010) Candidaemia associated with decreased in vitro fluconazole susceptibility: is *Candida*speciation predictive of the susceptibility pattern? *The Journal of Antimicrobial Chemotherapy* 65(7): 1460–1465.

Pais P, Galocha M, Takahashi-Nakaguchi A, et al. (2022) Multiple genome analysis of *Candida glabrata* clinical isolates renders new insights into genetic diversity and drug resistance determinants. *Microbial Cell* 9(11): 174–189.

Paluchowska P, Tokarczyk M, Bogusz B, et al. (2014) Molecular epidemiology of *Candida albicans* and *Candida glabrata* strains isolated from intensive care unit patients in Poland. *Memorias do Instituto Oswaldo Cruz* 109(4): 436–441.

Pankhurst CL (2013) Candidiasis (oropharyngeal). *BMJ Clinical Evidence* 2013: 1304.

Pappas PG, Kauffman CA, Andes D, et al. (2009) Clinical practice guidelines for the management of candidiasis: 2009 update by the Infectious Diseases Society of America. *Clinical Infectious Diseases* 48(5): 503–535.

Pappas PG, Kauffman CA, Andes DR, et al. (2016) Clinical Practice Guideline for the Management of Candidiasis: 2016 Update by the Infectious Diseases Society of America. *Clinical infectious diseases* 62(4): e1-50.

Pappas PG, Lionakis MS, Arendrup MC, et al. (2018) Invasive candidiasis. *Nature Reviews Disease Primers* 4(1): 18026.

Parazzini F, Di Cintio E, Chiantera V, et al. (2000) Determinants of different *Candida* species infections of the genital tract in women. *European Journal of Obstetrics, Gynecology, and Reproductive Biology* 93(2): 141–145.

Patel M (2022) Oral Cavity and *Candida albicans*: Colonisation to the Development of Infection. *Pathogens* 11(3): 335.

Pegueroles C, Mixão V, Carreté L, et al. (2020) HaploTypo: a variant-calling pipeline for phased genomes. *Bioinformatics* 36(8): 2569–2571.

Pemberton TJ (2006) Identification and comparative analysis of sixteen fungal peptidyl-prolyl cis/trans isomerase repertoires. *BMC Genomics* 7(1): 244.

Peter J, De Chiara M, Friedrich A, et al. (2018) Genome evolution across 1,011 *Saccharomyces cerevisiae* isolates. *Nature* 556(7701): 339–344.

Peterson SW, Demczuk W, Martin I, et al. (2023) Identification of bacterial and fungal pathogens directly from clinical blood cultures using whole genome sequencing. *Genomics* 115(2): 110580.

Pfaller MA, Messer SA, Hollis RJ, et al. (2009) Variation in susceptibility of bloodstream isolates of *Candida glabrata* to fluconazole according to patient age and geographic location in the United States in 2001 to 2007. *Journal of Clinical Microbiology* 47(10): 3185–3190.

Pfaller MA, Moet GJ, Messer SA, et al. (2011) Geographic variations in species distribution and echinocandin and azole antifungal resistance rates among *Candida* bloodstream infection isolates: report from the SENTRY Antimicrobial Surveillance Program (2008 to 2009). *Journal of Clinical Microbiology* 49(1): 396–399.

Pfaller MA, Castanheira M, Lockhart SR, et al. (2012) Frequency of decreased susceptibility and resistance to echinocandins among fluconazole-resistant bloodstream isolates of *Candida glabrata*. *Journal of Clinical Microbiology* 50(4): 1199–1203.

Pfaller MA, Rhomberg PR, Messer SA, et al. (2015) Isavuconazole, micafungin, and 8 comparator antifungal agents' susceptibility profiles for common and uncommon opportunistic fungi collected in 2013: temporal analysis of antifungal drug resistance using CLSI species-specific clinical breakpoints and proposed epidemiological cutoff values. *Diagnostic Microbiology and Infectious Disease* 82(4): 303–313.

Pfaller MA, Diekema DJ, Turnidge JD, et al. (2019) Twenty Years of the SENTRY Antifungal Surveillance Program: Results for *Candida* Species From 1997–2016. *Open Forum Infectious Diseases* 6(Supplement_1): S79–S94.

Pham LTT, Pharkjaksu S, Chongtrakool P, et al. (2019) A Predominance of Clade 17 *Candida albicans* Isolated From Hemocultures in a Tertiary Care Hospital in Thailand. *Frontiers in Microbiology* 10: 1194.

Phongpaichit S, Mackenzie DW and Fraser C (1987) Strain differentiation of *Candida albicans* by morphotyping. *Epidemiology and Infection* 99(2): 421–428.

Poirier S, Auger P, Joly J, et al. (1990) Interest of biotyping *Candida albicans* in chronic vulvovaginitis. *Mycoses* 33(1): 24–28.

Polonelli L, Archibusacci C, Sestito M, et al. (1983) Killer system: a simple method for differentiating *Candida albicans* strains. *Journal of Clinical Microbiology* 17(5): 774–780.

Polonelli L, Castagnola M, Rossetti DV, et al. (1985) Use of killer toxins for computer-aided differentiation of *Candida albicans* strains. *Mycopathologia* 91(3): 175–179.

Poplin R, Ruano-Rubio V, DePristo MA, et al. (2018) Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*. Available at: https://www.biorxiv.org/content/10.1101/201178v3 (accessed 15 August 2022).

Poulain D, Hopwood V and Vernes A (1985) Antigenic variability of *Candida albicans*. *Critical Reviews in Microbiology* 12(3): 223–270.

Price MN, Dehal PS and Arkin AP (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PloS One* 5(3): e9490.

Puius YA and Scully B (2007) Treatment of *Candida albicans* pericarditis in a heart transplant patient. *Transplant Infectious Disease* 9(3): 229–232.

Pujol C, Joly S, Nolan B, et al. (1999) Microevolutionary changes in *Candida albicans* identified by the complex *Ca3* fingerprinting probe involve insertions and deletions of the full-length repetitive sequence RPS at specific genomic sites. *Microbiology* 145(10): 2635–2646.

Pujol C, Pfaller M and Soll DR (2002) *Ca3* Fingerprinting of *Candida albicans* Bloodstream Isolates from the United States, Canada, South America, and Europe Reveals a European Clade. *Journal of Clinical Microbiology* 40(8): 2729.

Pujol C, Pfaller MA and Soll DR (2004) Flucytosine resistance is restricted to a single genetic clade of *Candida albicans*. *Antimicrobial Agents and Chemotherapy* 48(1): 262–266.

Purcell S, Neale B, Todd-Brown K, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* 81(3): 559–575.

Qu W-M, Liang N, Wu Z-K, et al. (2020) Minimum sample sizes for invasion genomics: Empirical investigation in an invasive whitefly. *Ecology and Evolution* 10(1): 38–49.

Quindós G, Fernández-Rodríguez M, Burgos A, et al. (1992) Colony morphotype on Sabouraud-triphenyltetrazolium agar: a simple and inexpensive method for *Candida* subspecies discrimination. *Journal of Clinical Microbiology* 30(10): 2748–2752.

Quindós G, Lipperheide V, Barturen B, et al. (1996) A new method of antibiotyping yeasts for subspecies discrimination and distribution in human clinical specimens. *European Journal of Epidemiology* 12(1): 55–62.

Ragon M, Wirth T, Hollandt F, et al. (2008) A new perspective on Listeria monocytogenes evolution. *PLoS Pathogens* 4(9): e1000146.

Rathod SD and Buffler PA (2014) Highly-cited estimates of the cumulative incidence and recurrence of vulvovaginal candidiasis are inadequately documented. *BMC Women's Health* 14(1): 43.

Rex J and Clinical (2009) Method for antifungal disk diffusion susceptibility testing of yeasts: Approved guideline.

Rhodes J, Abdolrasouli A, Dunne K, et al. (2022) Population genomics confirms acquisition of drug-resistant Aspergillus fumigatus infection by humans from the environment. *Nature Microbiology* 7(5): 663–674.

Richardson (2005) Changing patterns and trends in systemic fungal infections. *The Journal of Antimicrobial Chemotherapy.* 56(suppl_1):511.

Richter SS, Galask RP, Messer SA, et al. (2005) Antifungal susceptibilities of *Candida* species causing vulvovaginitis and epidemiology of recurrent cases. *Journal of Clinical Microbiology* 43(5): 2155–2162.

Roberts GD (1988) Medical mycology: The pathogenic fungi and the pathogenic Actinomycetes. *Mayo Clinic proceedings* 63(10): 1061–1062.

Roche CM, Loros JJ, McCluskey K, et al. (2014) *Neurospora crassa*: looking back and looking forward at a model microbe. *American Journal of Botany* 101(12): 2022–2035.

Rodrigues CF, Rodrigues ME and Henriques M (2019) *Candida* sp. Infections in Patients with Diabetes Mellitus. *Journal of Clinical Medicine* 8(1): 76.

Rodrigues de Miranda L (1979) *Clavispora*, a new yeast genus of the Saccharomycetales. *Antonie van Leeuwenhoek* 45(3): 479–483.

Roetzer A, Gratz N, Kovarik P, et al. (2010) Autophagy supports *Candida glabrata* survival during phagocytosis. *Cellular Microbiology* 12(2): 199–216.

Roetzer A, Klopf E, Gratz N, et al. (2011) Regulation of *Candida glabrata* oxidative stress resistance is adapted to host environment. *FEBS Letters* 585(2): 319–327.

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

Román MC and Linares Sicilia MJ (1983) Preliminary investigation of *Candida albicans* biovars. *Journal of clinical microbiology* 18(2): 430–431.

Romeo O, Tietz H-J and Criseo G (2013) *Candida africana*: Is It a Fungal Pathogen? *Current Fungal Infection Reports* 7(3): 192–197.

Ropars J, Maufrais C, Diogo D, et al. (2018) Gene flow contributes to diversification of the major fungal pathogen *Candida albicans*. *Nature Communications* 9(1): 2253.

Rosenberg SM (2011) Stress-induced loss of heterozygosity in *Candida*: a possible missing link in the ability to evolve. *mBio*. 2(5): 10-1128.

Rosentul DC, Delsing CE, Jaeger M, et al. (2014) Gene polymorphisms in pattern recognition receptors and susceptibility to idiopathic recurrent vulvovaginal candidiasis. *Frontiers in Microbiology* 5: 483.

Ruhnke M (2006) Epidemiology of *Candida albicans* infections and role of non-Candida-albicans yeasts. *Current Drug Targets* 7(4): 495–504.

Rustchenko E (2007) Chromosome instability in *Candida albicans*: Chromosome instability in *Candida albicans*. *FEMS Yeast Research* 7(1): 2–11.

Sah SK, Hayes JJ and Rustchenko E (2021) The role of aneuploidy in the emergence of echinocandin resistance in human fungal pathogen *Candida albicans*. *PLoS Pathogens* 17(5): e1009564.

Sala A, Ardizzoni A, Spaggiari L, et al. (2023) A new phenotype in Candida-epithelial cell interaction distinguishes colonization- versus vulvovaginal candidiasis-associated strains. *mBio* 14(2): e0010723.

Salama OE and Gerstein AC (2022) Differential Response of *Candida* Species Morphologies and Isolates to Fluconazole and Boric Acid. *Antimicrobial Agents and Chemotherapy* 66(5): e0240621.

Salazar SB, Pinheiro MJF, Sotti-Novais D, et al. (2022) Disclosing azole resistance mechanisms in resistant *Candida glabrata* strains encoding wild-type or gain-of-function CgPDR1 alleles through comparative genomics and transcriptomics. *G3* 12(7): jkac110.

Sampaio P, Gusmão L, Alves C, et al. (2003) Highly polymorphic microsatellite for identification of *Candida albicans* strains. *Journal of Clinical Microbiology* 41(2): 552–557.

Santos MAS, Gomes AC, Santos MC, et al. (2011) The genetic code of the fungal CTG clade. *Comptes Rendus Biologies* 334(8–9): 607–611.

Sautour M, Lemaître J-P, Ranjard L, et al. (2021) Detection and survival of *Candida albicans* in soils. *Environmental DNA* 3(6): 1093–1101.

Sayers EW, Bolton EE, Brister JR, et al. (2022) Database resources of the national center for biotechnology information. *Nucleic Acids Research* 50(D1): D20–D26.

Schaller M, Borelli C, Korting HC, et al. (2005) Hydrolytic enzymes as virulence factors of *Candida albicans*. *Mycoses* 48(6): 365–377.

Schliep KP (2011) phangorn: phylogenetic analysis in R. *Bioinformatics* 27(4): 592–593.

Schmid J, Herd S, Hunter PR, et al. (1999) Evidence for a general-purpose genotype in *Candida albicans*, highly prevalent in multiple geographical regions, patient types and types of infection. *Microbiology* 145(Pt 9): 2405–2413.

Schoch CL, Sung G-H, López-Giráldez F, 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.

Schröppel K, Rotman M, Galask R, et al. (1994) Evolution and replacement of *Candida albicans* strains during recurrent vaginitis demonstrated by DNA fingerprinting. *Journal of Clinical Microbiology* 32(11): 2646–2654.

Seelig MS (1966) The role of antibiotics in the pathogenesis of *Candida* infections. *The American Journal of Medicine* 40(6): 887–917.

Selmecki AM, Dulmage K, Cowen LE, et al. (2009) Acquisition of aneuploidy provides increased fitness during the evolution of antifungal drug resistance. *PLoS Genetics* 5(10): e1000705.

Shao T-Y, Ang WXG, Jiang TT, et al. (2019) Commensal *Candida albicans* positively calibrates systemic Th17 immunological responses. *Cell Host & Microbe* 25(3): 404-417.e6.

Sherrard J, Donders G, White D, et al. (2011) European (IUSTI/WHO) guideline on the management of vaginal discharge, 2011. *International Journal of STD & AIDS* 22(8): 421–429.

Shi W-M, Mei X-Y, Gao F, et al. (2007) Analysis of genital *Candida albicans* infection by rapid microsatellite markers genotyping. *Chinese medical journal* 120(11): 975–980.

Shi X-Y, Yang Y-P, Zhang Y, et al. (2015) Molecular identification and antifungal susceptibility of 186 *Candida*isolates from vulvovaginal candidiasis in southern China. *Journal of Medical Microbiology* 64(Pt 4): 390–393.

Siscar-Lewin S, Gabaldón T, Aldejohann AM, et al. (2021) Transient mitochondria dysfunction confers fungal cross-resistance against phagocytic killing and fluconazole. *mBio* 12(3): e0112821.

Sitterlé E, Maufrais C, Sertour N, et al. (2019) Within-Host Genomic Diversity of *Candida albicans* in Healthy Carriers. *Scientific reports* 9(1): 2563.

Sitterlé E, Coste AT, Obadia T, et al. (2020) Large-scale genome mining allows identification of neutral polymorphisms and novel resistance mutations in genes involved in *Candida albicans* resistance to azoles and echinocandins. *The Journal of Antimicrobial Chemotherapy* 75(4): 835–848.

Skrzypek MS, Binkley J, Binkley G, et al. (2017) The *Candida* Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic Acids Research* 45(D1): D592–D596.

Smith AC, Morran LT and Hickman MA (2022) Host Defense Mechanisms Induce Genome Instability Leading to Rapid Evolution in an Opportunistic Fungal Pathogen. *Infection and Immunity* 90(2): e0032821.

Smith DJ, Gold JAW, Williams SL, et al. (2025) An update on fungal disease outbreaks of public health concern. *Infectious Disease Clinics of North America* 39(1): 23–40.

Sobel JD (1986) Recurrent vulvovaginal candidiasis. A prospective study of the efficacy of maintenance ketoconazole therapy. *The New England Journal of Medicine* 315(23): 1455–1458.

Sobel JD (2003) Management of patients with recurrent vulvovaginal candidiasis. *Drugs* 63(11): 1059–1066.

Sobel JD (2007) Vulvovaginal candidosis. *Lancet* 369(9577): 1961–1971.

Sobel JD (2016) Recurrent vulvovaginal candidiasis. *American journal of Obstetrics and Gynecology* 214(1): 15–21.

Soll DR (2000) The ins and outs of DNA fingerprinting the infectious fungi. *Clinical Microbiology Reviews* 13(2): 332–370.

Soll DR and Daniels KJ (2016) Plasticity of *Candida albicans* biofilms. *Microbiology and Molecular Biology Reviews* 80(3): 565–595.

Soll DR and Pujol C (2003) *Candida albicans* clades. *FEMS Immunology and Medical Microbiology* 39(1): 1–7.

Song N, Kan S, Pang Q, et al. (2022) A prospective study on vulvovaginal candidiasis: multicentre molecular epidemiology of pathogenic yeasts in China. *Journal of the European Academy of Dermatology and Venereology* 36(4): 566–572.

Spinillo A, Nicola S, Colonna L, et al. (1994) Frequency and significance of drug resistance in vulvovaginal candidiasis. *Gynecologic and obstetric investigation* 38(2): 130–133.

Sprenger M, Hartung TS, Allert S, et al. (2020) Fungal biotin homeostasis is essential for immune evasion after macrophage phagocytosis and virulence. *Cellular Microbiology* 22(7): e13197.

Srikantha T, Lachke SA and Soll DR (2003) Three mating type-like loci in *Candida glabrata*. *Eukaryotic Cell* 2(2): 328–340.

Staib P and Morschhäuser J (2007) Chlamydospore formation in *Candida albicans* and *Candida*dubliniensis--an enigmatic developmental programme. *Mycoses* 50(1): 1–12.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9): 1312–1313.

Stappers MHT and Brown GD (2017) Host immune responses during infections with *Candida albicans*. In: *Candida albicans: Cellular and Molecular Biology*: 145–183.

Steensels J, Gallone B and Verstrepen KJ (2021) Interspecific hybridization as a driver of fungal evolution and adaptation. *Nature Reviews. Microbiology* 19(8): 485–500.

Stiller RL, Bennett JE, Scholer HJ, et al. (1982) Susceptibility to 5-fluorocytosine and prevalence of serotype in 402 *Candida albicans* isolates from the United States. *Antimicrobial Agents and Chemotherapy* 22(3): 482–487.

Stone W, Jones B-L, Wilsenach J, et al. (2012) External ecological niche for *Candida albicans* within reducing, oxygen-limited zones of wetlands. *Applied and Environmental Microbiology* 78(7): 2443–2445.

Sudbery P, Gow N and Berman J (2004) The distinct morphogenic states of *Candida albicans*. *Trends in Microbiology* 12(7): 317–324.

Sui Y, Qi L, Wu J-K, et al. (2020) Genome-wide mapping of spontaneous genetic alterations in diploid yeast cells. *Proceedings of the National Academy of Sciences* 117(45): 28191–28200.

Sundstrom P, Balish E and Allen CM (2002) Essential role of the *Candida albicans* transglutaminase substrate, hyphal wall protein 1, in lethal oroesophageal candidiasis in immunodeficient mice. *The Journal of Infectious Diseases* 185(4): 521–530.

Szarvas J, Rebelo AR, Bortolaia V, et al. (2021) Danish Whole-Genome-Sequenced *Candida albicans* and *Candida glabrata* Samples Fit into Globally Prevalent Clades. *Journal of Fungi:* 7(11).

Tabara H, Sarkissian M, Kelly WG, et al. (1999) The rde-1 gene, RNA interference, and transposon silencing in C. elegans. *Cell* 99(2): 123–132.

Taj-Aldeen SJ, Kolecka A, Boesten R, et al. (2014) Epidemiology of candidemia in Qatar, the Middle East: performance of MALDI-TOF MS for the identification of *Candida* species, species distribution, outcome, and susceptibility pattern. *Infection* 42(2): 393–404.

Talazadeh F, Ghorbanpoor M and Shahriyari A (2022) Candidiasis in birds (Galliformes, Anseriformes, Psittaciformes, Passeriformes, and Columbiformes): A focus on antifungal susceptibility pattern of *Candida albicans* and non-albicans isolates in avian clinical specimens. *Topics in Companion Animal Medicine* 46(100598): 100598.

Tasić S, Tasić N, Tasić A, et al. (2002) Recurrent genital candidosis of women; consequence of reinfection of relapse. *Medical Biology*. 9: 217-222.

Tavanti A, Gow NAR, Senesi S, et al. (2003) Optimization and validation of multilocus sequence typing for *Candida albicans*. *Journal of clinical microbiology* 41(8): 3765–3776.

Thompson DS, Carlisle PL and Kadosh D (2011) Coevolution of morphology and virulence in *Candida*species. *Eukaryotic cell* 10(9): 1173–1182.

Tian J-Y, Yang Y-G, Chen S, et al. (2021) Genetic diversity and molecular epidemiology of *Candida albicans* from vulvovaginal candidiasis patients. *Infection, genetics and evolution. Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 92: 104893.

Timmermans B, De Las Peñas A, Castaño I, et al. (2018) Adhesins in *Candida glabrata*. *Journal of Fungi* 4(2): 60.

Todd RT and Selmecki A (2020) Expandable and reversible copy number amplification drives rapid adaptation to antifungal drugs. *eLife* 9: e58349.

Todd RT, Wikoff TD, Forche A, et al. (2019) Genome plasticity in *Candida albicans* is driven by long repeat sequences. *eLife* 8: e45954.

Turgut M, Challa S and Akhaddar A (2019) *Fungal Infections of the Central Nervous System: Pathogens, Diagnosis, and Management*. Springer.

Vale-Silva L, Beaudoing E, Tran VDT, et al. (2017) Comparative Genomics of Two Sequential *Candida glabrata* Clinical Isolates. *G3* 7(8): 2413–2426.

Van der Auwera GA, Carneiro MO, Hartl C, et al. (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics* 43(1): 11-10.

van Rhijn N and Rhodes J (2025) Evolution of antifungal resistance in the environment. *Nature microbiology* 10(8): 1804–1815.

van Schalkwyk J, Yudin MH and INFECTIOUS DISEASE COMMITTEE (2015) Vulvovaginitis: screening for and management of trichomoniasis, vulvovaginal candidiasis, and bacterial vaginosis. *Journal of Obstetrics and Gynaecology Canada* 37(3): 266–274.

Vande Zande P, Zhou X and Selmecki A (2023) The dynamic fungal genome: Polyploidy, aneuploidy and copy number variation in response to stress. *Annual Review of Microbiology* 77: 341–361.

Venkateswarlu K, Taylor M, Manning NJ, et al. (1997) Fluconazole tolerance in clinical isolates of *Cryptococcus neoformans*. *Antimicrobial agents and chemotherapy* 41(4): 748–751.

Vila T, Sultan AS, Montelongo-Jauregui D, et al. (2020) Oral Candidiasis: A Disease of Opportunity. *Journal of Fungi:* 6(1): 15.

Wainwright M, Wickramasinghe NC, Narlikar JV, et al. (2003) Microorganisms cultured from stratospheric air samples obtained at 41 km. *FEMS Microbiology Letters* 218(1): 161–165.

Ward TL, Dominguez-Bello MG, Heisel T, et al. (2018) Development of the human mycobiome over the first month of life and across body sites. *mSystems* 3(3): 10-1128.

Whelan WL, Kirsch DR, Kwon-Chung KJ, et al. (1990) *Candida albicans* in patients with the acquired immunodeficiency syndrome: absence of a novel of hypervirulent strain. *The journal of Infectious Diseases* 162(2): 513–518.

White TC (1997) The presence of an R467K amino acid substitution and loss of allelic variation correlate with an azole-resistant lanosterol 14alpha demethylase in *Candida albicans*. *Antimicrobial Agents and Chemotherapy* 41(7): 1488–1494.

WHO (2022) *WHO fungal priority pathogens list to guide research, development and public health action*. 25 October. Available at: https://apps-who-int.uml.idm.oclc.org/iris/rest/bitstreams/1474282/retrieve.

Williamson MI, Samaranayake LP and MacFarlane TW (1986) Biotypes of oral *Candida albicans* and *Candida*tropicalis isolates. *Journal of medical and veterinary mycology: bi-monthly publication of the International Society for Human and Animal Mycology* 24(1): 81–84.

Williamson MI, Samaranayake LP and MacFarlane TW (1987) A new simple method for biotyping *Candida albicans*. *Microbios* 51(208–209): 159–167.

Xu J, Schwartz K, Bartoces M, et al. (2008) Effect of antibiotics on vulvovaginal candidiasis: a MetroNet study. *Journal of the American Board of Family Medicine:* 21(4): 261–268.

Xu YY and Samaranayake LP (1995) Oral *Candida albicans* biotypes in Chinese patients with and without oral candidosis. *Archives of Oral Biology* 40(6): 577–579.

Xu Z, Green B, Benoit N, et al. (2021) Cell wall protein variation, break-induced replication, and subtelomere dynamics in *Candida glabrata*. *Molecular Microbiology* 116(1): 260–276.

Yamaguchi MU, Rampazzo R de CP, Yamada-Ogatta SF, et al. (2007) Yeasts and filamentous fungi in bottled mineral water and tap water from municipal supplies. *Brazilian Archives of Biology and Technology* 50(1): 1–9.

Yang F, Todd RT, Selmecki A, et al. (2021) The fitness costs and benefits of trisomy of each *Candida albicans* chromosome. *Genetics* 218(2): iyab056.

Yano J, Sobel JD, Nyirjesy P, et al. (2019) Current patient perspectives of vulvovaginal candidiasis: incidence, symptoms, management and post-treatment outcomes. *BMC Women's Health* 19(1): 48.

Yates AD, Allen J, Amode RM, et al. (2022) Ensembl Genomes 2022: an expanding genome resource for non-vertebrates. *Nucleic Acids Research* 50(D1): D996–D1003.

Zhang G, Chen Y, Chen J, et al. (2024) Association of multilocus sequence typing, MSH2 gene mutations, and antifungal resistance in *Candida glabrata*: implications for clinical outcomes in Chinese hospitals. *Annals of Clinical Microbiology and Antimicrobials* 23(1): 100.

Zhang J-Y, Liu J-H, Liu F-D, et al. (2014) Vulvovaginal candidiasis: species distribution, fluconazole resistance and drug efflux pump gene overexpression. *Mycoses* 57(10): 584–591.

Zheng L, Kelly CJ and Colgan SP (2015) Physiologic hypoxia and oxygen homeostasis in the healthy intestine. A Review in the Theme: Cellular Responses to Hypoxia. *American Journal of Physiology* 309(6): C350-60.

Zheng Q, Liu J, Qin J, et al. (2022) Ploidy variation and spontaneous haploid-diploid switching of *Candida glabrata* clinical isolates. *mSphere* 7(4): e0026022.

Zhou Y, Cheng L, Lei YL, et al. (2021) The interactions between *Candida albicans* and mucosal immunity. *Frontiers in microbiology* 12: 652725.

Zhu Y, Fang C, Shi Y, et al. (2022) *Candida albicans* multilocus sequence typing clade I contributes to the clinical phenotype of vulvovaginal candidiasis patients. *Frontiers in Medicine* 9: 837536.

Zordan RE, Galgoczy DJ and Johnson AD (2006) Epigenetic properties of white-opaque switching in *Candida albicans* are based on a self-sustaining transcriptional feedback loop. *Proceedings of the National Academy of Sciences* 103(34): 12807–12812.

Zuber J, Sah SK, Mathews DH, et al. (2023) Genome-wide DNA changes acquired by *Candida albicans* caspofungin-adapted mutants. *Microorganisms* 11(8): 1870.

Zupan J and Raspor P (2008) Quantitative agar-invasion assay. *Journal of Microbiological Methods* 73(2): 100–104.