

Linking geography, isolation source, and genomic diversity in a global *Candida albicans* phylogeny

Abdul-Rahman Adamu Bukari

University of Manitoba

Dana A. Opulente

J. F. Crow Institute for the Study of Evolution, University of Wisconsin-Madison

Christopher Todd Hittinger

J. F. Crow Institute for the Study of Evolution, University of Wisconsin-Madison

Christoph d'Enfert

Université Paris Cité

Christina A. Cuomo

Brown University

Aleeza C. Gerstein

Aleeza.Gerstein@umanitoba.ca

University of Manitoba

Article

Keywords:

Posted Date: March 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8970909/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: Competing interest reported. Christina A. Cuomo serves as an Associate Editor of this journal and had no role in the peer review or decision to publish this manuscript. All other authors declare no competing interests.

Abstract

Candida albicans is a common commensal species in multiple sites in the human microbiome that can also be an opportunistic pathogen across the body. Previous phylogenomic analyses have identified major clades, but these studies often relied on imprecise genomic methods or unphased genomes, limited geographic and ecological sampling, and a phylogenetic resolution strategy that has not been universally standardized within the research community. Here, we address these gaps by reconstructing a whole-genome phylogeny to examine how geography and site of isolation contribute to phylogenetic structure in *C. albicans*. We analyzed phased genomes from 938 global isolates acquired from diverse clinical and ecological contexts, including soil, and applied an agnostic, threshold-based clustering approach to systematically define cluster boundaries. In addition, we examined genomic features such as aneuploidy, the distribution of mating-type locus (*MTL*), genome-wide heterozygosity, and RNA interference (RNAi) disruption. Our analyses preserved the previously defined major clusters while identifying six novel clusters, predominantly composed of highly admixed Asian isolates. Although geographic origin and isolation source were each significantly associated with cluster, these associations were confounded because isolates from specific regions were disproportionately derived from particular sources, preventing attribution of the observed clustering to either factor. Over 95% of the isolates were heterozygous at the *MTL*, although homozygous forms were enriched in some clusters. Analysis of the *AGO1* PAZ domain revealed both known and novel RNAi variants, predominantly in a heterozygous state. Aneuploidy was present in 8% of isolates, spread across the phylogeny. Intra-host analysis of isolates from 95 people revealed predominantly clonal colonization, though fourteen of the individuals harboured multiple genetic clusters. This study refines the phylogenetic structure of *C. albicans*, demonstrating how genomic features such as aneuploidy, heterozygosity, *MTL* composition, and RNAi disruption vary across isolates and provide insights into genomic plasticity in this species.

Introduction

Candida albicans is a common human fungal opportunistic pathogen responsible for a wide range of conditions, from superficial to life-threatening^{1–3}. *C. albicans* is also commonly found as a commensal species in many of the same body sites where it causes disease^{4–6}. The switch to pathogenicity is known to occur when the local microbiota is disrupted, tissue barriers are compromised, or the immune defenses are weakened; thus, the same strains that exist as commensal are often implicated in disease⁷. *C. albicans* has also been isolated from animals that are closely associated with humans⁸ as well as human-associated environments (e.g., food spoilage). A small number of environmental isolates from sand, soil, and tree bark have been acquired^{9–12}. Two out of three *C. albicans* isolates from oak tree bark were closely related to human isolates, the third was not assigned to an existing cluster¹¹. The transmission dynamics of *C. albicans* are not fully resolved. Humans are typically colonized at birth from their mother^{13,14}, and available evidence suggests that commensal strains generally remain stable within individuals in early life¹⁵. However, detailed longitudinal data on within-host strain diversity and microevolution remain limited¹⁶.

The *C. albicans* diploid genome is about 14.5 Mb and encodes ~ 6,000 genes¹⁷. *C. albicans* is predominantly asexual and hence its genome is at least partially released from the karyotypic constraint imposed by frequent meiosis. In addition to "standard" single-nucleotide variation (SNVs) and small insertions and deletions (INDELS), short- and long-tract copy number changes, loss of heterozygosity (LOH), and chromosomal aneuploidy are also frequently observed in lab experiments and among clinical isolates^{18–21}. 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^{7,22}. Epigenetic variation, enabling rapid acclimation, potentially 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 of Ago1 (argonaute), the central RNAi component. However, a recent assessment of 295 additional isolates found that only eight isolates carry the missense variant²³. Of these, seven variants are heterozygous and likely retain RNAi activity, while only one additional isolate was homozygous for the inactive SC5314 variant.

The 182-isolate whole-genome sequencing (WGS) phylogeny widely used for *C. albicans* was published in 2018; we refer to this tree as the "existing WGS tree"²¹. The existing WGS tree is comprised of 17 distinct genetic clusters, including 12 numbered according to the numbering of clusters (clades) previously identified by multilocus sequence typing^{24,25}. The cutoff value used for cluster designation analysis from the MLST study was itself partially based on clusters identified by an earlier DNA-fingerprinting method using the moderately repetitive Ca3 probe^{26,27}. 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^{11,21,28}. Five additional clusters were identified in the WGS data and assigned alphabetical designations (A-E); each contained fewer than 10 isolates. Support for the existing WGS cluster delineation was found from NGSAdmix analysis (with $k = 13$), which identified the same clusters. NGSAdmix additionally indicated extensive admixture in the ten isolates that were labelled as singletons, and evidence for recombination in common ancestors leading to new WGS cluster A (between clusters 3 and D) and cluster B (between clusters 2 and E). A recent admixture analysis of MLST data of over 5,000 isolates identified extensive admixture across most geographic populations (with $k = 2$), consistent with frequent gene flow²⁹.

In some fungal species, isolates within the same cluster tend to have similar geographic origins, virulence, and/or resistance patterns (*Candidozyma auris*³⁰, *Saccharomyces cerevisiae*³¹, *Aspergillus fumigatus*^{32,33}). The pattern in *C. albicans* seems more nuanced, though in some studies, isolates from the same geographic region cluster together. By example, recent MLST analyses of vulvovaginal isolates from northern China revealed a novel cluster of 92 *C. albicans* isolates³⁴, while a study of isolates from Thailand found that ~ 25% (13/46) clustered together in MLST cluster 17³⁵ though this cluster also contains isolates from other continents²⁴). It stands to reason that any potential geographical specificity

of clusters 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^{24,34,36}. *C. albicans* may be particularly susceptible to this due to its high colonization prevalence in healthy hosts compared to other fungal opportunistic pathogens such as *Cryptococcus neoformans* and *C. auris*.

Many studies have sought to determine whether there is a link between cluster and phenotypic traits of clinical interest. Cluster 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 clusters^{37–39}. Cluster 1 isolates are also the most likely to be resistant to terbinafine compared to clusters 2, 3, 4 and 11 isolates³⁹, yet no difference was observed for a panel of seven other common antifungal drugs. Isolates in cluster 2 have significantly lower levels of acid phosphatase activity than cluster 1 and 3 isolates⁴⁰. Significant differences among isolates in different clusters (1, 2, 3, 4) in relation to intergenic tandem repeat 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 identified⁴⁰. Blood isolates from cluster 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 cluster association was observed in terms of biofilm formation, growth in bovine serum albumin, growth in different temperatures and adherence to plastic catheter⁴⁰. A significantly higher proportion of isolates from cluster 1 are associated with superficial infections and commensal carriage compared to isolates from other clusters, and cluster 1 is repeatedly the largest cluster, regardless of the phylogenetic method used^{24,36}. It has been hypothesized that cluster 1 isolates may have an enhanced ability to evade host defenses due to the possession of a 985 bp *HpaII* 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 study⁴¹.

To improve phylogenetic resolution, address the aforementioned limitations in cluster designation, and assess potential signatures of geography and sites of isolation, we reconstructed a new short-read whole-genome sequencing (WGS) phylogeny for *C. albicans*. We used data from 1,178 isolates, including 1,130 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 cluster 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.

Results

1,178 *C. albicans* isolates with WGS data obtained from 26 countries across five continents were included in the study. Geographic and site of isolation distributions were very uneven (Figure 1, Table S1). North America (520 isolates) and Asia (410 isolates) were the best-represented continents, with many fewer isolates from Europe (179), Africa (51), and South America (10); no isolates came from Australia. At the country level, sampling was even sparser; only 11 countries had at least ten isolates. Unfortunately, the majority of isolates from each country came from a single isolation source; only eight countries had isolates from multiple sources. Approximately one-third of all isolates were from the bloodstream, and another third were oral. The “environmental” isolates included 19 from food spoilage (all from France), two from birds (both from France), and ten from plants or soil (three from oak trees in the UK, and seven isolates were newly sequenced for this study in the United States, with samples collected from soil (2 isolates) and plant-associated sources (5 isolates). This uneven distribution was 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.

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 phased tree was constructed with 2,252,881 SNPs, approximately twice the number used in the unphased tree (1,113,404 SNPs), while the MLST tree relied on 197 SNPs. The majority of nodes in all generated trees have bootstrap support above 80%, a metric which can thus raise our confidence in tree topologies of intraspecific phylogenies based on WGS data (Figure 2). The normalized Robinson-Foulds (nRF) distance (a measure of topological dissimilarity between trees) was 0.655 when the phased and unphased WGS trees were compared. 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 influences the tree topology, which can also be seen by visually comparing the trees (Figure 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.

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 (1,178) 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 S1). The majority of removed isolates came from North America (n=217). The 35 isolates that were identified as cluster 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).

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

To delineate clusters for the updated tree, we sought to adopt a statistically supported approach that could be easily replicated on an extended or different strain set. We implemented seven TreeCluster strategies that determine phylogenetic tree topologies by clustering tip sequences with different optimization functions and distance thresholds. Given that isolate designations have been relatively consistent over time through different sequencing technologies, *a priori* we were expecting to find regions of threshold space that would yield approximately 17 clusters (the number from the existing tree) with substantial overlap in cluster assignment with the existing WGS phylogeny. Five strategies showed no threshold regions of cluster number stability, with the majority of values at either a single cluster or a very high (>50) numbers of clusters (Figure 3A). Two strategies, however, yielded broad regions of parameter space where the same number of clusters was identified (Figure 3A). The default, "max clade" strategy, identified 20 clusters from threshold parameter space values between 0.012-0.014, while the "single linkage" strategy (which has previously been used in HIV research) identified 24 clusters over values 0.047-0.055, excluding 0.051. However, the 24 clusters predicted by the single linkage strategy combined approximately 3/4 of the isolates into a single cluster, with all additional clusters identified among the isolates closest to the root of the tree (Figure 3B).

When we mapped the max clade cluster designations onto the final phylogeny, it was visually clear that many clusters matched the previous WGS tree well (Figure 4). Fourteen clusters from the previous WGS phylogeny were retained. Both clusters with previous evidence of gene flow from the admixture analysis in the existing WGS tree were merged into one of their ancestral parent clusters (cluster B was nested into cluster 2, and cluster A was nested into cluster D). We assigned new numeric designations to the three existing clusters that had previously been assigned letters (cluster A/D was renamed as cluster 19, cluster C was renamed to cluster 20, and cluster E was renamed as 21). These new clade numbers start at 19, following from the 18 numbered clusters identified in the existing phylogeny. In addition, six novel clusters were identified that had not been identified in the existing phylogeny; these were numbered 22-27 based on their phylogenetic location in a counter-clockwise manner (Figure 4).

Branch colours correspond to cluster designations as defined in the study. The cluster numbers are denoted on the clusters, and asterisks are used to show new cluster labels. The red clusters are labels given to existing alphabetically designated clusters. Bootstrap support values $\geq 80\%$ are indicated by ribbons on the branches. Two concentric rings surround the phylogeny: the inner ring represents the continent of origin for each isolate, while the outer ring denotes the anatomical source of isolation. This visualization highlights both the phylogenetic structure and the geographic and ecological diversity across clusters.

We compared the new WGS phylogeny to an MLST phylogeny; that is, we extracted the MLST sequence information from the phylogenetically informative isolates and generated a new phylogenetic tree. Although many isolates from the same WGS cluster grouped together in the MLST tree, there were also many cases of discordance, where isolates from different WGS clusters grouped together, or isolates from the same WGS cluster were apart (Figure S3). Six MLST clusters were absent in the existing WGS tree (clusters 5, 6, 7, 14, 15, and 17). When we overlay our WGS cluster designations with all of the available MLST sequence information (i.e., from our isolates as well as those from pubMLST⁴², isolates from the absent MLST clades did not group with any of the new WGS clusters (Figure S4). As we lack WGS sequence information for these historical isolates, as in the existing WGS tree, we did not reassign those cluster numbers.

A previous admixture analysis of isolates included in the existing phylogeny identified two clusters that contained isolates with multiple ancestries at $k = 13$ ²¹. In addition, a visual inspection of the existing admixture plot shows that the 10 isolates previously classified as singletons were also highly admixed. However, admixture signals observed in singleton or sparsely represented groups should be interpreted with caution, as populations represented by fewer than five individuals are known to yield unreliable ancestry estimates⁴³. When we reran an admixture analysis on the updated phylogeny from $k = 6$ to $k = 20$, a much higher proportion of isolates appear to be admixed than previously identified (Figure 5, Figure S5). At $k = 20$ (i.e., the same number of ancestries as statistically identified clusters), 60 % ($n = 501$) of the isolates were identified as single ancestry, while 20% ($n = 140$) of the isolates were highly admixed (from 4-18 ancestries). Interestingly, at either $k = 13$ (Figure 5A), or $k = 20$ (Figure 5B), the majority of clusters that were previously identified to contain almost entirely single ancestry isolates either continued to show this (clusters 2, 3, 4, 9, 18, 21), or contained sub-clusters, a large one with single ancestry isolates and a smaller one with mixed ancestry isolates (clusters 8, 11, 12, 22, 23)(Figure 5C). The exceptions were the isolates in cluster 1, which at $k = 20$ (but not $k = 13$), had either two or three ancestries, isolates in cluster 10 (which were all highly admixed at all k values), and isolates in cluster 20 (formerly cluster C), which contained several blocks of highly admixed isolates. Whether cluster 1 contains predominantly single- or mixed-ancestry isolates depends heavily on k : most k values below 13 classify the isolates as single ancestry, whereas higher k values classify them as mixed ancestry (Figure S5). Cluster 19, which contains the isolates previously classified as mixed ancestry in cluster A, is similar to cluster 1, with isolate classification (single or from two ancestries) depending on the k value chosen (Figure S5). The six new clusters were split between those with highly admixed ancestry isolate

subclusters (clusters 22, 23, 25) and those containing entirely highly admixed isolates (clusters 24, 26, and 27).

Geography and site of isolation are nested together

The distribution of isolates in the full phylogeny was largely consistent with the existing phylogeny. Cluster 1 contained approximately a quarter of all isolates, while three clusters (10, 16, 27) had fewer than ten isolates. cluster designation, isolation source (Figure S6 and S7), and geographic origin (Figure S8) are not independent of each other. This is seen visually on the map (Figure 1) and the overlapping colour blocks in the outer rings of the phylogeny (e.g., North American oral isolates exhibit several clean blocks in multiple clusters; Figure 4). There was a significant statistical association between all three pairwise factor comparisons, though the association between isolation source and continent had the largest effect size (pairwise Chi-square tests with 100,000 Monte Carlo simulations; source $\times$ continent: $\chi^2 = 1064$, $p < 0.0001$, Cramér's $V = 0.49$; cluster $\times$ source: $\chi^2 = 433$, $p < 0.0001$, $V = 0.23$; cluster $\times$ continent: $\chi^2 = 473$, $p < 0.0001$, $V = 0.33$). Isolates from the blood were overrepresented in Asia relative to North America (and, to a lesser extent, Europe), whereas

oral isolates were similarly overrepresented in North America relative to Asia (Figure S9). Interestingly, Africa and South America are both over-represented for vaginal isolates, while Europe was over-represented for environmental and urogenital isolates (Figure S9). These results suggest that while there is some significant phylogenetic structure based on geography and isolation source, these factors are currently largely confounded and heavily influenced by research focus.

Manitoba as a case study to remove geography

To examine the phylogenetic relationships among isolates from different sources with the potential confounding factor of geography removed, we analyzed a set of 83 isolates collected from a hospital microbiology lab in Manitoba (this entails all *C. albicans* isolates collected in 2012 and 2018). The isolates fell into nine different clusters, eight of which were represented in both years; there was no significant difference in cluster composition between the two years ($\chi^2 = 6.64$, $p = 0.6534$, Figure S10). The isolates were obtained from nine isolation sources; notably, no significant association was observed between cluster and isolation source ($\chi^2 = 66.90$, $p = 0.3814$), and isolates from different sources (and different years) often grouped right beside each other. This, as previously shown, highlights that there is no clear pattern between cluster and isolation source, and that migration across the globe appears relatively common among *C. albicans* genotypes. However, this picture remains incomplete, as a signature of geographic enrichment is present in many clusters. Indeed, the novel clusters we identify in our strain set, relative to the existing WGS tree, are largely composed of Asian isolates. Combining the global and local (Manitoba) analyses suggests that geography, rather than isolation source, has more of an impact on shaping phylogenetic relationships.

Karyotypic variation

The full isolate set was examined for aneuploidies and copy number variations (CNVs; Figure S11). 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 6A). Four of these isolates were identified as triploid (3N), four as tetraploid (4N), and the 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 than their size would have predicted, suggesting there might be stronger selection against extra copies of these chromosomes. Blood isolates were most likely to have a CNV or aneuploidy of chromosomes 6, 7 and R; oral isolates were most likely to have karyotypic variants on chromosomes 4 or 5; while karyotypic variants were relatively equally distributed among chromosomes for vaginal isolates. 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 6B). 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 6B).

Karyotypic variant isolates were observed throughout the phylogeny and were typically not clustered together (Figure S11). At least one isolate from each isolation source was observed (Figure 6C), with the number proportional to the total number of isolates from each source. The three most common sources exhibited different patterns, which hints at a potential beneficial association between selection in these different niches (blood versus oral) selecting for different chromosomal variants. However, caution in interpretation is needed given the sampling biases. The bloodstream isolates are from multiple locations and authored publications, while the oral isolates all come from a single study on North American isolates focused on drug resistance⁴⁴.

MTL locus analyses

The distribution and variation of mating-type loci (*MTL*) were examined in the phylogenetically informative set of isolates (Table S1, Figure S11). 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 ($a/a/a$: 1 isolate; $a/a/\alpha$: 4

isolates; a/a/a: 0.3%; α/a/a: 1 isolate). Homozygous *MTL* isolates were from every continent and all major isolation sources. A statistically significant but modest association were observed between *MTL* genotype and continent (Fisher's Exact Test with Monte Carlo simulation, $p = 0.016$) and *MTL* genotype and cluster ($p = 0.012$), while there was no association between

MTL genotype and isolation source ($p = 0.075$). That isolation source was not significant suggests that neutral processes likely drive variation in the distribution of *MTL* genotypes.

Heterozygosity analyses

Heterozygosity was assessed across the 744 isolates with genomic information for $\geq 85\%$ of sites. The mean genome-wide heterozygosity in *C. albicans* was 0.0065 ± 0.001 (SD), corresponding to approximately 6.5 SNPs/kb, and ranged from 0.0029 to 0.014 (2.9–14 SNPs/kb). To look at the potential effect of different factors on genome-wide heterozygosity, we did a three-factor ANOVA with isolation source, cluster, and geography. There was a significant interaction between isolation source and cluster 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: $F_{9, 615} = 17.4$, $p < 0.0001$, cluster: $F_{20, 615} = 14.5$, $p < 0.0001$, isolation source $\times$ cluster: $F_{85, 615} = 1.7$, $p < 0.0001$, Figure 7A). Post-hoc comparisons identified that blood isolates had significantly higher heterozygosity than isolates from many other sites, though the isolate distributions across different sites were highly overlapping (see Figure 6A for statistical results). Analysis of heterozygosity across the eight chromosomes revealed variability in mean values, with chromosome 6 exhibiting the highest average (0.00814 ± 0.00324 ; ~ 8.14 SNPs/kb) and chromosome R the lowest (0.00572 ± 0.00152 ; ~ 5.72 SNPs/kb (Table S2).

Although the majority of isolates have a relatively similar level of genome-wide heterozygosity, there were several outlier isolates that were significantly lower or higher (Figure 7B). The seven high heterozygosity isolates are all from China, except for one from the US and they were predominantly isolated from the blood (though one is oral, and one is gastrointestinal). Although these seven isolates are admixed at $K=13$ and $K=20$, highly admixed ancestry profiles do not uniformly correspond to the highest levels of heterozygosity, indicating that admixture alone is insufficient to explain the observed heterozygosity. By contrast, the three low heterozygosity isolates are all single ancestry isolates from three different sources (oral, vaginal, blood) and three different clusters (2, 3, 20) from North America. To determine whether specific regions of the genome were causative, for each isolate we divided the genome into 5kb bins and counted the number of heterozygous positions in each. There were many small LOH events across the isolates. The LOH regions encompassed regions common to all isolates within clusters, later events that are common to only some members of a cluster, and isolate-specific events (Figure S12). Notably, the high heterozygosity isolates contained heterozygous positions throughout the genome (Figure 7C). However, terminal end of chromosome R was consistently depleted

for heterozygosity across most isolates, including the highly heterozygous isolates. This suggests that there was either an ancestral LOH event prior to the most recent common ancestor of the species, or that there have been several recurrent LOH events extending from the rDNA toward the telomere, a pattern that was previously observed in *S. cerevisiae*⁴⁵ and thought to be driven by elevated recombination at the rDNA promoted by a conflict between transcription and replication. (Figure 8, Figure S12). 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 common regions of both high and low heterozygosity, 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 < 1 heterozygous base /kb. This identified 119 bins; the majority (n = 78) are at the terminal right end of chromosome R. Three additional regions are located on chromosomes 1 (one bin), 2 (one bin), 3 (twelve bins), 4 (one bin) and 5 (two bins, Figure 8). Many named genes with known or predicted functions that could be the target of purifying selection are present (Table S3). There were overall fewer bins with elevated heterozygosity, defined as those with >100 heterozygous bases on average (i.e., heterozygosity > 0.1 /kb); they were found in 10 distinct genomic regions distributed across five chromosomes (Figure 8). After excluding bins that include annotated repeat sequences, we identified 12 unique protein-coding genes within these variable regions. The genes included *HAL21*, *CAN1*, *RIM9*, *KAR5*, and *MSY1*, along with several encoding proteins with unknown functions (Table S3). Further work is required to validate whether there is a fitness benefit to increased heterozygosity or disruptive selection (e.g., favouring variation among isolates in different ecological contexts) in these genes.

Low-heterozygosity regions (defined as < 5 SNPs per window) are shown as orange circles (upper panel), while high-heterozygosity regions (defined as > 100 SNPs per window) are indicated by purple circles (lower panel).

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 mutated in the SC5314 reference strain. We identified 43 unique SNPs in the PAZ domain among the phylogenetically informative set of 908 *C. albicans* (excluding cluster 13 isolates): 17 were synonymous mutations, while 26 were non-synonymous (Figure 9).

The RNAi-active consensus PAZ sequence was homozygous in 796 isolates (88%), which were distributed across the phylogenetic tree (Figure 9A 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 cluster 1 isolates, while a heterozygous version of this allele was found in an additional 19 cluster 1 isolates and one cluster 4 isolate (Figure 9B). The var2 homozygous variant was found in a single cluster 19 isolate, with the heterozygous form most commonly found in the same cluster. Other variants, although heterozygous, showed cluster-specific enrichment: var3 was predominant in a subcluster of cluster 20, while var5 was mainly linked to cluster 9.

Intra-individual analysis

Nine studies included more than one isolate collected from the same individual or from a healthy mother-infant dyad (Table S4). Nearly all individuals were from the United States (75) and Canada (12), four were from Spain, and one each from Morocco, France, Brazil and Tunisia. The largest study included oral isolates from 59 mother-infant dyads in 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 with either oral and fecal or rectal and vaginal site pairs) 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 clusters: 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 clusters (oral and fecal isolates from a healthy individual). More studies on intra-individual isolates (including isolates from the same site taken at the same time) are required to fully characterize intra-population diversity under different contexts.

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 4). One soil isolate, however, was a singleton that did not cluster with any other isolates and was an outgroup to the most rest of the *C. albicans* phylogeny. The two European isolates from starlings were positioned adjacent to the outgroup in a small cluster (cluster 16) with only three other isolates (a urogenital isolate from Europe, an oral isolate from the UK and a blood isolate from China). This suggests they fall into one of the earliest-diverging lineages among the sampled isolates.

The average heterozygosity of environmental isolates did not differ significantly from other sources (Figure 6A), nor were they more likely to be aneuploid, 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 near the outgroup, 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.

Discussion

The distribution of *C. albicans* isolates that have had their genomes sequenced with short-read technologies 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. Similar to the disparity among WGS isolates, among the ~ 6000 isolates submitted to the pubMLST database (as of July 30, 2025), only 197 are from Africa and 88 from Oceania^{42,46}. The lack of genomic studies from these locations, which show some of the highest diversity in other taxa, highlights a critical gap in capturing local and global *C. albicans* intraspecific 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 specific disease presentations. For example, a large proportion of isolates come from studies on candidemia (particularly from Asia^{47,48}), oral candidiasis (particularly from North America^{15,44}) and commensal colonization^{15,49,50}. Vaginal isolates are the most evenly sampled across continents, consistent with this being the most common disease presentation. Environmental isolates are rare, and only a handful of isolates have been collected from genuinely natural environments such as soil¹² (and the seven new sequences) or tree bark¹¹, rather than clearly human-associated environments, such as food spoilage²¹. Despite their scarcity, the environmental isolates generally show high genomic similarity to isolates from human-associated isolates. Furthermore, the three oak tree isolates exhibited either similar or higher host-damage potential and intrinsic resistance to amphotericin B and fluconazole compared to clinical isolates, indicating that environmental *C. albicans* can harbour substantial pathogenic and drug-resistant traits⁵¹.

C. albicans has been shown to be transmitted through vertical (mother-to-child)^{13,14} and intrafamilial transmission⁵². Isolates from the same geographic location, obtained from different isolation sources, often cluster together, suggesting the potential for extrafamilial routes of transmission. It would be interesting to look at isolates over a longer period of time to determine whether there is strain turnover in later life stages. There is a pressing need for expanded sampling across multiple scales to better understand the factors that drive relatedness in the *C. albicans* phylogeny.

We used phased whole-genome data, together with a statistical tool, to construct a phylogeny and identify 20 clusters. The phylogenetic structure is largely consistent with the existing WGS tree²¹, while identifying six additional clusters. The isolates in the new clusters are predominantly from Asia. A recent study from China that added 369 isolates to the existing WGS tree, proposed 38 clusters, including 21 novel clusters⁴⁸. In our analyses (which also included these isolates), the novel clusters described by Gong et al. were either merged into existing clusters or were classified as one of the six novel clusters we described. From the historical MLST cluster nomenclature, six cluster numbers are unaccounted for in our phylogeny (clusters 5, 6, 7, 14, 15, 17). None of the novel WGS clusters correspond to these missing MLST clusters, as unfortunately, no WGS isolates were available from those clusters and isolates from these clusters should be prioritized for short-read sequencing in the future. We proposed an expanded cluster designation for new WGS clusters with numbers starting at 19, which included revising the alphabetic cluster nomenclature introduced by Ropars et al.²¹ to a fully numeric system. This approach standardizes clade nomenclature, minimizes confusion across studies, and conservatively maintains continuity with prior work. Under this framework, we recognize the potential for 27 identified clusters (20 from our WGS tree and up to seven from the MLST legacy isolates).

The lack of strong geographic or isolation source-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 set²¹. The specific chromosomes involved differed between blood and oral isolates (chromosomes 6, 7, and R for blood; 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 phenotype⁵³. However, studies have also shown that chromosome 7 trisomy can enhance colonization of both the gastrointestinal tract and the oral cavity in mouse models^{54,55,56}. 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* isolate⁵⁷, aneuploidy was also observed in isolate T118 during serial passage in fluconazole, yet did not directly contribute to enhanced drug resistance⁵⁸.

The distribution of mating-type locus (*MTL*) genotype was also relatively consistent with previous large-scale surveys. The homozygous mating-type loci were relatively rare (2.7%) and within the range of previously reported homozygosity rates of 2.2%^{21,59} and 3.2%^{21,59} (though Odds *et al.* observed a slightly higher prevalence of 8.5%²⁴). In all studies, including this one, the α/α genotype was more common than a/a , suggesting that the loss of the *MTLa* 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 (0.0066¹¹ and an analysis of 61 diverse isolates (0.0067⁶⁰ while higher than the average from Ropars et al. (0.0048^{21,59}. 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. Blood isolates, for example, displayed higher heterozygosity than gastrointestinal or oral isolates, potentially reflecting the immune-challenged and dynamic nature of the circulatory system. 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^{21,48,61}, yet these differences are potentially driven by biased sampling, with the confounding variable of isolation source nested into clade.

The majority of isolates (> 99%) maintained the RNAi-active consensus PAZ sequence in at least one copy, indicating that RNAi activity is likely preserved in the vast majority of *Candida albicans* isolates. However, we identified novel genetic alterations in the PAZ domain. A previous study examined 296 isolates and identified a single homozygous mutation (361 E/K) in the SC5314 reference strain and seven other isolates with RNAi deficiency²³. They also identified isolates containing heterozygous variants at positions 345P and 365I that retained RNAi activity. In our largest dataset, we identified ten isolates with homozygous variants: five carried the 361 E/K known RNAi-deficient variant, one was homozygous for 365 I/V, one was homozygous for 341 R/K, and three were homozygous for 346 T/N. The known 361 E/K isolates are all closely related in cluster 1, while the other homozygous variants are at disparate places in the phylogeny. Their limited spread implies a potential fitness cost or selective constraint associated with loss of RNAi function. As our study was predominantly *in silico*, we did not assess the potential 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.

Conclusion

This study refines the phylogenetic structure of *C. albicans* using phased whole-genome data, resolving both known and novel clusters 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*.

Methods

Acquisition of sequence data

The NCBI SRA database was searched in July 2024 for all deposited Illumina sequenced *C. albicans* whole genome sequences⁴⁶. Isolates from experimental evolution studies were excluded. Additionally, only samples with $\geq 80\%$ of reads mapping to the reference genome, breadth of coverage $\geq 80\%$ and mean depth of coverage $\geq 30\times$ were considered (following human WGS analyses^{62,63}). In total, 1088 isolates from 22 publications were retained^{11,15,21,28,44,47–50,61,64–75}. In addition, we included fastq data generated by the Joint Genome Institute from seven newly sequenced environmental isolates; with samples collected from soil (2 isolates) and plant-associated sources (5 isolates); and 83 newly sequenced isolates acquired in 2012 and 2018 from the microbiology lab of the major hospital in Manitoba (Table S1) for a total of 1178 isolates ("complete isolate set"). The isolate collection included 361 isolates from 95 individuals with more than one isolate ("intrapopulation isolates"): 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 S1). When necessary to avoid potential bias, when intrapopulation isolates from the same body site clustered monophyletically, we randomly selected a single isolate from each monophyletic group for each individual.

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 described⁷⁶. DNA quality and concentration were assessed using the Thermo Scientific™ NanoDrop 2000 and Qubit® 2.0 Fluorometer (with the Invitrogen™ Qubit™ dsDNA BR Assay Kit), 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

Reads were inspected with FastQC v0.11.5

(<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and trimmed and filtered with Trimmomatic v0.36⁷⁷. Read mapping and variant calling were performed using HaploTypo v1.0.1⁷⁸ using the default parameters. Briefly, Haplotipo is run in four successive modules (see scripts in <https://github.com/Gabaldonlab/haplotipo>:) 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 (haplomap.py). Filtered reads were mapped with readgroups added onto the phased haplotypes (hapA and hapB) of the A21 SC5314 reference genome¹⁷ with BWA-MEM v0.7.18⁷⁹. 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 CollectAlignmentSummaryMetrics from picard v3.1.0 (<http://broadinstitute.github.io/picard>)

and consolidated across all samples with MultiQC⁸⁰. Average depth of coverage was assessed with samtools coverage⁸¹ of samtools v1.20. Bcftools v1.19⁸¹ was used for calling and filtering variants using the default parameters in Haplotipo. The VCFcorr_alleles.py script of Haplotipo 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).

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.*²¹ (i.e., the existing WGS phylogeny) as there exists cluster information for these isolates. We retained only one of the 35 cluster 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 haplotipo. 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 package⁸² in the R Programming language. 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.

WGS phylogeny construction from the complete isolate set

A new phylogeny was generated from 945 of "phylogenetically informative" isolates. This included 128 of the 361 intrapopulation isolates (as described above). A FASTA file for the reconstructed haplotype from each isolate was generated using the `haplomap.py` of `haplotyp` script. This process ensures that heterozygous positions are not disregarded or replaced by IUPAC ambiguity codes, as is observed in many pipelines⁸³. 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 (2.1.11)⁸⁴ 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 RAxML⁸⁵, within a significantly shorter time⁸⁶. The 37 *C. africana* ("cluster 13") isolates were used to root the tree^{87,88}. The resulting phylogeny was visualized and annotated with the Interactive Tree Of Life (iTOL, v5)⁸⁹.

MLST phylogenetic tree construction

Multilocus sequence typing (MLST) alleles were derived both from whole-genome sequencing (WGS) data and from publicly available MLST records. Extraction of MLST loci from WGS assemblies is not trivial, as allele boundaries can vary between isolates and may not be directly retrievable from draft genomes. To accurately recover MLST loci from WGS data, individual locus reference sequences were first aligned to each genome assembly to identify the precise genomic coordinates of each locus. These coordinates were then used to extract the corresponding sequences from all samples. Extracted loci were subsequently concatenated in a consistent order to generate full MLST profiles for each isolate. In addition, MLST sequences were downloaded from the PubMLST database to supplement the dataset. Duplicate isolates were identified and removed prior to phylogenetic analysis to avoid redundancy and over-representation of closely related samples. Concatenated MLST sequences were used to construct a phylogenetic tree representing the complete MLST dataset. The resulting tree was examined to confirm that distinct DSTs formed separate clades, verifying that sequence types were phylogenetically distinguishable within the dataset. The sequence types of the 83 newly sequenced Canadian isolates, as well as the seven environmental isolates from the USA, were determined using stringMLST⁹⁰ with default parameters. The *Candida albicans* MLST database was downloaded from PubMLST. Comparison with the PubMLST database showed that only three of the 83 Canadian isolates, as well as the seven environmental isolates from the USA, corresponded to previously described DSTs; the remaining isolates did not match any known DSTs.

Genetic cluster delineation

To delineate the clusters within the phylogeny, we ran TreeCluster⁹¹. TreeCluster uses several functions to agnostically identify clusters within phylogenetic trees. We selected methods that are optimized to identify clusters 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 genetic clusters. We then visually inspected each genetic cluster assignment on the phylogeny using iTOL and compared it to the existing WGS phylogeny.

Heterozygosity analyses

Sites with missing data are common in short-read WGS datasets and can substantially bias estimates of heterozygosity. We therefore excluded isolates with more than 15% missing genotype positions from the heterozygosity analyses, retaining 745 isolates from the phylogenetically informative isolate set. This filtering step was applied specifically to heterozygosity calculations because missing data disproportionately inflate or deflate per-isolate heterozygosity estimates, whereas phylogenetic inference is comparatively robust to moderate levels of missing data due to the large number of informative sites contributing to tree topology. For this reason, isolates exceeding the missingness threshold were retained in phylogenetic analyses but excluded from heterozygosity comparisons. Average heterozygosities were calculated and statistically compared by isolation site and cluster after verifying the homogeneity of variances using Levene’s test.

Genome-wide heterozygosity at each was calculated using PLINK v2.0⁹². Bcftools was used to call variants in the consensus mode to ensure all sites are considered using the A21 hapA reference. 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.

NGSAdmix

Admixture was estimated among isolates of the phylogenetically informative set. Genotype likelihoods were estimated directly from aligned sequencing reads using ANGSD⁹³. Major and minor alleles were inferred from the data, and SNPs were identified using a likelihood ratio test with a significance threshold of $p < 1 \times 10^{-6}$. Minor allele frequencies were estimated for all retained sites. Genotype likelihoods were

output in BEAGLE format for downstream population genetic analyses. All analyses were parallelized using 10 computational threads. Population structure was inferred using NGSadmixture, which estimates individual ancestry proportions from genotype likelihoods. Analyses were performed using genotype likelihoods in BEAGLE format. We evaluated a range of ancestral population numbers ($k = 6-20$), running each value of k independently. Loci with a minor allele frequency below 0.05 were excluded. Each NGSadmixture run was executed using 32 threads, and analyses were parallelized across K values using GNU Parallel. Admixture proportions were visualized in R.

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 S13.

MTL analyses

The mating type-like (*MTL*) locus 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 **MTLb** region (Ca22chr5B_C_albicans_SC5314:395642–396223 and 401608–402227). Loci with coverage < 0.5 were considered inactive or absent.

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.

Declarations

COI:

Christina A. Cuomo serves as an Associate Editor of this journal and had no role in the peer review or decision to publish this manuscript. All other authors declare no competing interests.

Competing Interests

Christina A. Cuomo serves as an Associate Editor of this journal and had no role in the peer review or decision to publish this manuscript. All other authors declare no competing interests.

Author Contribution

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.A.O contributed environmental isolates, C.C. hosted A-R.A.B at the Broad Institute and contributed conceptually to the bioinformatic analyses, C.d'E. contributed conceptually to 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. All authors provided suggestions on manuscript drafts and approved the final manuscript.

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 PRJNA1418157. All data and code required to reproduce all statistical analyses and visualizations (excepting the phylogenies) are available at https://github.com/microstatslab/Calbians_phylogenetics. Files exceeding 50 MB could not be uploaded due to repository size limits and are available from the authors upon request.

Acknowledgements

We thank Shared Health, Diagnostic Services Manitoba and Dr. Markus Stein for facilitating the acquisition of clinical strains from the Health Sciences Centre in Winnipeg, Manitoba. We thank the members of the MicroStats lab through the years for many helpful comments throughout this project. The computational research was enabled in large part by support from Ali Kerrache (Prairies Region) and by resources available through the Digital Research Alliance of Canada (<https://www.alliancecan.ca/>). This project was funded by the National Science and Engineering Research Council of Canada (RGPIN-2019-05867) and a MITACS Globalink award to A.C.G, A.-R.A.B. and C.A.C. A.C.G. acknowledges the support of the CIFAR Azrieli Global Scholars Program and start-up funding from the University of Manitoba. A.-R.A.B. was supported by an EvoFunPath (NSERC CREATE) fellowship. This project was also supported by the National Science Foundation under Grants No. DEB-2110403, the National Institute of Food and Agriculture, United States Department of Agriculture, Hatch project 7005101, and based upon work at the Great Lakes Bioenergy Research Center supported by the U.S. Department of Energy, Office of Science, Biological and Environmental Research under Award Number DE-SC0018409. The work conducted by the U.S. Department of Energy Joint Genome Institute (<https://ror.org/04xm1d337>), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. Work in the laboratory of CdE is supported by the Agence Nationale de la Recherche (ANR-10-LABX-62-IBEID). C.A.C. is supported by the National Institutes of Health, NIAID grant U19AI110818-011. The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

References

1. Ruhnke, M. Epidemiology of *Candida albicans* infections and role of non-*Candida-albicans* yeasts. *Curr. Drug Targets* **7**, 495–504 (2006).
2. Brown, G. D. *et al.* Hidden killers: human fungal infections. *Sci. Transl. Med.* **4**, 165rv13 (2012).
3. Friedman, D. Z. P. & Schwartz, I. S. Emerging Fungal Infections: New Patients, New Patterns, and New Pathogens. *J Fungi (Basel)* **5**, (2019).
4. Ghannoum, M. A. *et al.* Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. *PLoS Pathog.* **6**, e1000713 (2010).
5. Drell, T. *et al.* Characterization of the vaginal micro- and mycobiome in asymptomatic reproductive-age Estonian women. *PLoS One* **8**, e54379 (2013).
6. Kashem, S. W. & Kaplan, D. H. Skin Immunity to *Candida albicans*. *Trends Immunol.* **37**, 440–450 (2016).
7. d'Enfert, C. *et al.* The impact of the Fungus-Host-Microbiota interplay upon *Candida albicans* infections: current knowledge and new perspectives. *FEMS Microbiol. Rev.* **45**, (2021).
8. Talazadeh, F., Ghorbanpoor, M. & Shahriyari, A. 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. *Top. Companion Anim. Med.* **46**, 100598 (2022).

9. Maciel, N. O. *et al.* Occurrence, antifungal susceptibility, and virulence factors of opportunistic yeasts isolated from Brazilian beaches. *Mem. Inst. Oswaldo Cruz* **114**, e180566 (2019).
10. Hamlin, J. A. P., Dias, G. B., Bergman, C. M. & Bensasson, D. Phased Diploid Genome Assemblies for Three Strains of *Candida albicans* from Oak Trees. *G3* (2019) doi:10.1534/g3.119.400486.
11. Bensasson, D. *et al.* Diverse Lineages of *Candida albicans* Live on Old Oaks. *Genetics* **211**, 277–288 (2019).
12. Opulente, D. A. *et al.* Pathogenic budding yeasts isolated outside of clinical settings. *FEMS Yeast Res.* **19**, (2019).
13. Filippidi, A. *et al.* The effect of maternal flora on *Candida* colonisation in the neonate. *Mycoses* **57**, 43–48 (2014).
14. Caramalac, D. A. *et al.* *Candida* isolated from vaginal mucosa of mothers and oral mucosa of neonates: occurrence and biotypes concordance. *Pediatr. Infect. Dis. J.* **26**, 553–557 (2007).
15. Alkhars, N., Al Jallad, N., Wu, T. T. & Xiao, J. Multilocus sequence typing of *Candida albicans* oral isolates reveals high genetic relatedness of mother-child dyads in early life. *PLoS One* **19**, e0290938 (2024).
16. Ward, T. L. *et al.* Development of the human mycobiome over the first month of life and across body sites. *mSystems* **3**, (2018).
17. Muzzey, D., Schwartz, K., Weissman, J. S. & Sherlock, G. Assembly of a phased diploid *Candida albicans* genome facilitates allele-specific measurements and provides a simple model for repeat and indel structure. *Genome Biol.* **14**, R97 (2013).
18. Forche, A. *et al.* Rapid Phenotypic and Genotypic Diversification After Exposure to the Oral Host Niche in *Candida albicans*. *Genetics* **209**, 725–741 (2018).
19. Smith, A. C., Morran, L. T. & Hickman, M. A. Host Defense Mechanisms Induce Genome Instability Leading to Rapid Evolution in an Opportunistic Fungal Pathogen. *Infect. Immun.* **90**, e0032821 (2022).
20. Sui, Y. *et al.* Genome-wide mapping of spontaneous genetic alterations in diploid yeast cells. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 28191–28200 (2020).
21. Ropars, J. *et al.* Gene flow contributes to diversification of the major fungal pathogen *Candida albicans*. *Nat. Commun.* **9**, 2253 (2018).
22. Mayer, F. L., Wilson, D. & Hube, B. *Candida albicans* pathogenicity mechanisms. *Virulence* **4**, 119–128 (2013).
23. Iracane, E. *et al.* Identification of an active RNAi pathway in *Candida albicans*. *Proc. Natl. Acad. Sci. U. S. A.* **121**, e2315926121 (2024).
24. Odds, F. C. *et al.* Molecular phylogenetics of *Candida albicans*. *Eukaryot. Cell* **6**, 1041–1052 (2007).
25. Shin, J. H. *et al.* Genetic diversity among Korean *Candida albicans* bloodstream isolates: assessment by multilocus sequence typing and restriction endonuclease analysis of genomic DNA by use of BssHII. *J. Clin. Microbiol.* **49**, 2572–2577 (2011).

26. Soll, D. R. & Pujol, C. *Candida albicans* clades. *FEMS Immunol. Med. Microbiol.* **39**, 1–7 (2003).
27. Pujol, C., Pfaller, M. & Soll, D. R. 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**, 2729 (2002).
28. Szarvas, J. *et al.* Danish Whole-Genome-Sequenced *Candida albicans* and *Candida glabrata* Samples Fit into Globally Prevalent Clades. *J Fungi (Basel)* **7**, (2021).
29. Dalmieda, J. & Xu, J. Global population genetics and evolutionary dynamics of *Candida albicans*. *Can. J. Microbiol.* (2026) doi:10.1139/cjm-2025-0248.
30. Chow, N. A. *et al.* Tracing the Evolutionary History and Global Expansion of *Candida auris* Using Population Genomic Analyses. *MBio* **11**, (2020).
31. Loegler, V., Friedrich, A. & Schacherer, J. Overview of the *Saccharomyces cerevisiae* population structure through the lens of 3,034 genomes. *G3 (Bethesda)* **14**, jkae245 (2024).
32. Rhodes, J. *et al.* Population genomics confirms acquisition of drug-resistant *Aspergillus fumigatus* infection by humans from the environment. *Nat. Microbiol.* **7**, 663–674 (2022).
33. He, X. *et al.* Genomic diversity of the pathogenic fungus *Aspergillus fumigatus* in Japan reveals the complex genomic basis of azole resistance. *Commun. Biol.* **7**, 274 (2024).
34. Song, N. *et al.* A prospective study on vulvovaginal candidiasis: multicentre molecular epidemiology of pathogenic yeasts in China. *J. Eur. Acad. Dermatol. Venereol.* **36**, 566–572 (2022).
35. Pham, L. T. T., Pharkjaksu, S., Chongtrakool, P., Suwannakarn, K. & Ngamskulrungroj, P. A Predominance of Clade 17 *Candida albicans* Isolated From Hemocultures in a Tertiary Care Hospital in Thailand. *Front. Microbiol.* **10**, 1194 (2019).
36. Odds, F. C. & Jacobsen, M. D. Multilocus sequence typing of pathogenic *Candida* species. *Eukaryot. Cell* **7**, 1075–1084 (2008).
37. Pujol, C., Pfaller, M. A. & Soll, D. R. Flucytosine resistance is restricted to a single genetic clade of *Candida albicans*. *Antimicrob. Agents Chemother.* **48**, 262–266 (2004).
38. Dodgson, A. R., Dodgson, K. J., Pujol, C., Pfaller, M. A. & Soll, D. R. Clade-specific flucytosine resistance is due to a single nucleotide change in the FUR1 gene of *Candida albicans*. *Antimicrob. Agents Chemother.* **48**, 2223–2227 (2004).
39. Odds, F. C. In *Candida albicans*, resistance to flucytosine and terbinafine is linked to MAT locus homozygosity and multilocus sequence typing clade 1. *FEMS Yeast Res.* **9**, 1091–1101 (2009).
40. MacCallum, D. M. *et al.* Property differences among the four major *Candida albicans* strain clades. *Eukaryot. Cell* **8**, 373–387 (2009).
41. Giblin, L. *et al.* A DNA polymorphism specific to *Candida albicans* strains exceptionally successful as human pathogens. *Gene* **272**, 157–164 (2001).
42. Bournoux, M.-E. *et al.* Collaborative consensus for optimized multilocus sequence typing of *Candida albicans*. *J. Clin. Microbiol.* **41**, 5265–5266 (2003).

43. Jørsboe, E., Hanghøj, K. & Albrechtsen, A. fastNGSadmix: admixture proportions and principal component analysis of a single NGS sample. *Bioinformatics* **33**, 3148–3150 (2017).
44. Ford, C. B. *et al.* The evolution of drug resistance in clinical isolates of *Candida albicans*. *Elife* **4**, e00662 (2015).
45. Lindstrom, D. L., Leverich, C. K., Henderson, K. A. & Gottschling, D. E. Replicative age induces mitotic recombination in the ribosomal RNA gene cluster of *Saccharomyces cerevisiae*. *PLoS Genet.* **7**, e1002015 (2011).
46. Sayers, E. W. *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022).
47. Chew, K. L., Achik, R., Osman, N. H., Octavia, S. & Teo, J. W. P. Genomic epidemiology of human candidaemia isolates in a tertiary hospital. *Microb Genom* **9**, (2023).
48. Gong, J. *et al.* Emergence of antifungal resistant subclades in the global predominant phylogenetic population of *Candida albicans*. *Microbiol. Spectr.* **11**, e0380722 (2023).
49. Anderson, F. M. *et al.* *Candida albicans* selection for human commensalism results in substantial within-host diversity without decreasing fitness for invasive disease. *PLoS Biol.* **21**, e3001822 (2023).
50. Sitterlé, E. *et al.* Within-Host Genomic Diversity of *Candida albicans* in Healthy Carriers. *Sci. Rep.* **9**, 2563 (2019).
51. Lange, T. *et al.* “Pour some sugar on me”-Environmental *Candida albicans* isolates and the evolution of increased pathogenicity and antifungal resistance through sugar adaptation. *PLoS Pathog.* **21**, e1013542 (2025).
52. Bougnoux, M.-E. *et al.* Multilocus sequence typing reveals intrafamilial transmission and microevolutions of *Candida albicans* isolates from the human digestive tract. *J. Clin. Microbiol.* **44**, 1810–1820 (2006).
53. Forche, A. *et al.* Selection of *Candida albicans* trisomy during oropharyngeal infection results in a commensal-like phenotype. *PLoS Genet.* **15**, e1008137 (2019).
54. Kakade, P., Sircaik, S., Maufrais, C., Ene, I. V. & Bennett, R. J. Aneuploidy and gene dosage regulate filamentation and host colonization by *Candida albicans*. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2218163120 (2023).
55. Ene, I. V. *et al.* Global analysis of mutations driving microevolution of a heterozygous diploid fungal pathogen. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E8688–E8697 (2018).
56. Mishra, A. *et al.* Strain background interacts with chromosome 7 aneuploidy to determine commensal and virulence phenotypes in *Candida albicans*. *PLoS Genet.* **21**, e1011650 (2025).
57. Anderson, M. Z., Saha, A., Haseeb, A. & Bennett, R. J. A chromosome 4 trisomy contributes to increased fluconazole resistance in a clinical isolate of *Candida albicans*. *Microbiology* **163**, 856–865 (2017).

58. Selmecki, A. M., Dulmage, K., Cowen, L. E., Anderson, J. B. & Berman, J. Acquisition of aneuploidy provides increased fitness during the evolution of antifungal drug resistance. *PLoS Genet.* **5**, e1000705 (2009).
59. Lockhart, S. R. *et al.* In *Candida albicans*, white-opaque switchers are homozygous for mating type. *Genetics* **162**, 737–745 (2002).
60. Mixão, V. & Gabaldón, T. Genomic evidence for a hybrid origin of the yeast opportunistic pathogen *Candida albicans*. *BMC Biol.* **18**, 48 (2020).
61. Hirakawa, M. P. *et al.* Genetic and phenotypic intra-species variation in *Candida albicans*. *Genome Res.* **25**, 413–425 (2015).
62. Mardis, E. R. Applying next-generation sequencing to pancreatic cancer treatment. *Nat. Rev. Gastroenterol. Hepatol.* **9**, 477–486 (2012).
63. Ajay, S. S., Parker, S. C. J., Abaan, H. O., Fajardo, K. V. F. & Margulies, E. H. Accurate and comprehensive sequencing of personal genomes. *Genome Res.* **21**, 1498–1505 (2011).
64. Cavalieri, D. *et al.* Genomic and Phenotypic Variation in Morphogenetic Networks of Two *Candida albicans* Isolates Subtends Their Different Pathogenic Potential. *Front. Immunol.* **8**, 1997 (2017).
65. McTaggart, L. R., Cabrera, A., Cronin, K. & Kus, J. V. 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. *Antimicrob. Agents Chemother.* **64**, (2020).
66. Chew, K. L., Octavia, S., Jureen, R., Lin, R. T. P. & Teo, J. W. P. Targeted amplification and MinION nanopore sequencing of key azole and echinocandin resistance determinants of clinically relevant *Candida* spp. from blood culture bottles. *Lett. Appl. Microbiol.* **73**, 286–293 (2021).
67. Gnaïen, M. *et al.* A gain-of-function mutation in zinc cluster transcription factor Rob1 drives *Candida albicans* adaptive growth in the cystic fibrosis lung environment. *PLoS Pathog.* **20**, e1012154 (2024).
68. Li, X. V. *et al.* Immune regulation by fungal strain diversity in inflammatory bowel disease. *Nature* **603**, 672–678 (2022).
69. Mohammadi, S., Leduc, A., Charette, S. J., Barbeau, J. & Vincent, A. T. Amino acid substitutions in specific proteins correlate with farnesol unresponsiveness in *Candida albicans*. *BMC Genomics* **24**, 93 (2023).
70. Peterson, S. W. *et al.* Identification of bacterial and fungal pathogens directly from clinical blood cultures using whole genome sequencing. *Genomics* **115**, 110580 (2023).
71. Sitterlé, E. *et al.* Large-scale genome mining allows identification of neutral polymorphisms and novel resistance mutations in genes involved in *Candida albicans* resistance to azoles and echinocandins. *J. Antimicrob. Chemother.* **75**, 835–848 (2020).
72. Zuber, J., Sah, S. K., Mathews, D. H. & Rustchenko, E. Genome-wide DNA changes acquired by *Candida albicans* caspofungin-adapted mutants. *Microorganisms* **11**, 1870 (2023).

73. Guinea, J. *et al.* Whole genome sequencing confirms *Candida albicans* and *Candida parapsilosis* microsatellite sporadic and persistent clones causing outbreaks of candidemia in neonates. *Med. Mycol.* **60**, (2021).
74. Cuomo, C. A. *et al.* Genome sequence for *Candida albicans* clinical oral isolate 529L. *Microbiol. Resour. Announc.* **8**, (2019).
75. Adamu Bukari, A.-R. *et al.* Migration and standing variation in vaginal and rectal yeast populations in recurrent vulvovaginal candidiasis. *mSystems* **10**, e0015725 (2025).
76. Kukurudz, R. J. *et al.* Acquisition of cross-azole tolerance and aneuploidy in *Candida albicans* strains evolved to posaconazole. *G3* (2022).
77. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
78. Pegueroles, C., Mixão, V., Carreté, L., Molina, M. & Gabaldón, T. HaploTypo: a variant-calling pipeline for phased genomes. *Bioinformatics* **36**, 2569–2571 (2020).
79. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
80. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
81. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).
82. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
83. Ortiz, E. M. *Vcf2phyliip v2.0: Convert a VCF Matrix into Several Matrix Formats for Phylogenetic Analysis.* (2019). doi:10.5281/zenodo.2540861.
84. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
85. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
86. Liu, K., Linder, C. R. & Warnow, T. RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS One* **6**, e27731 (2011).
87. Romeo, O., Tietz, H.-J. & Criseo, G. *Candida africana*: Is It a Fungal Pathogen? *Curr. Fungal Infect. Rep.* **7**, 192–197 (2013).
88. Mixão, V., Saus, E., Boekhout, T. & Gabaldón, T. Extreme diversification driven by parallel events of massive loss of heterozygosity in the hybrid lineage of *Candida albicans*. *Genetics* **217**, (2021).
89. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
90. Gupta, A., Jordan, I. K. & Rishishwar, L. stringMLST: a fast k-mer based tool for multilocus sequence typing. *Bioinformatics* **33**, 119–121 (2017).
91. Balaban, M., Moshiri, N., Mai, U., Jia, X. & Mirarab, S. TreeCluster: Clustering biological sequences using phylogenetic trees. *PLoS One* **14**, e0221068 (2019).

92. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

93. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics* **15**, 356 (2014).

Figures

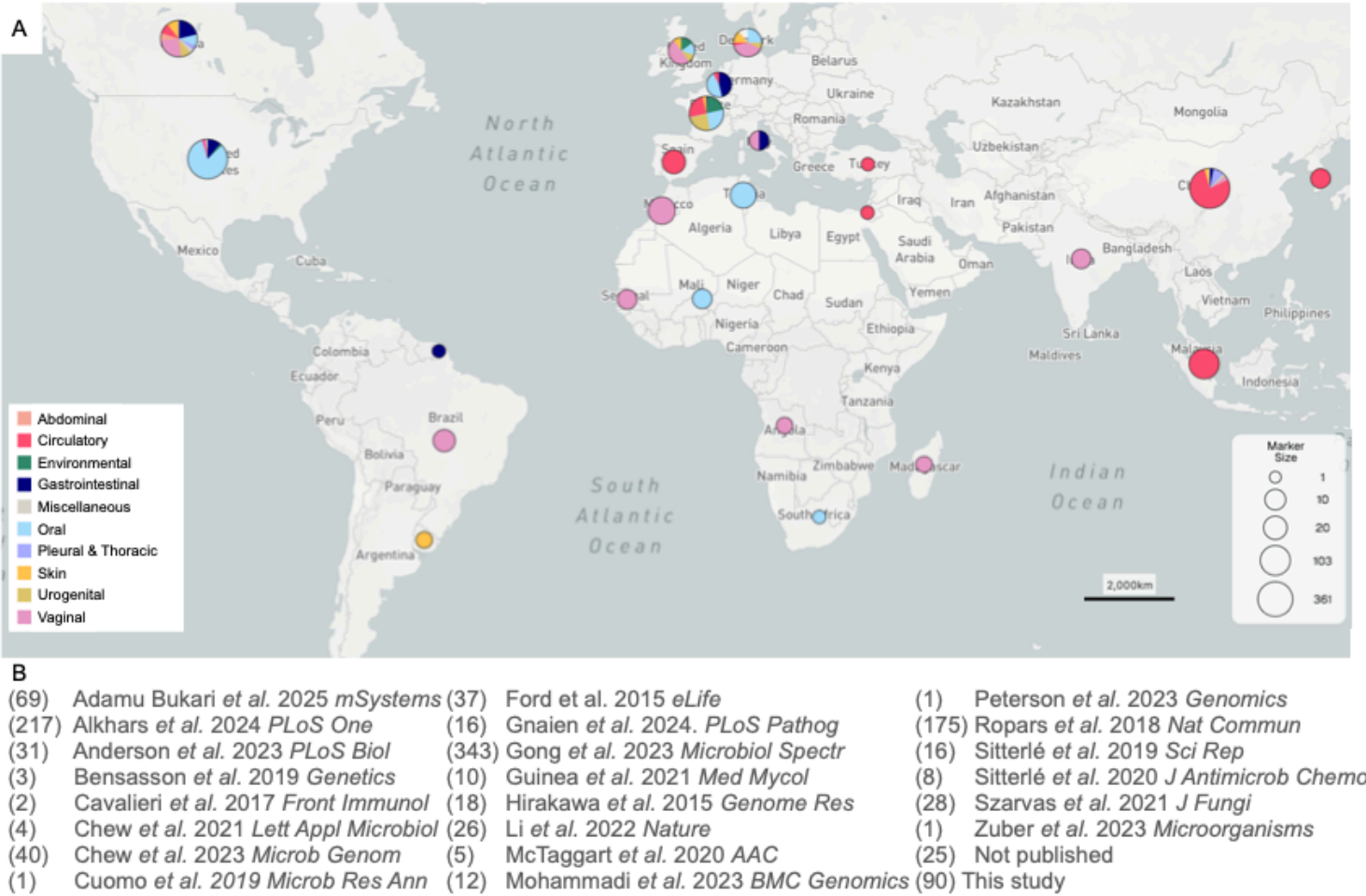


Figure 1

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 cluster 13 (*Candida africana*) isolates. The image was generated using Microreact. B) Publication source of the isolates used. The number of isolates from each manuscript is provided in brackets.

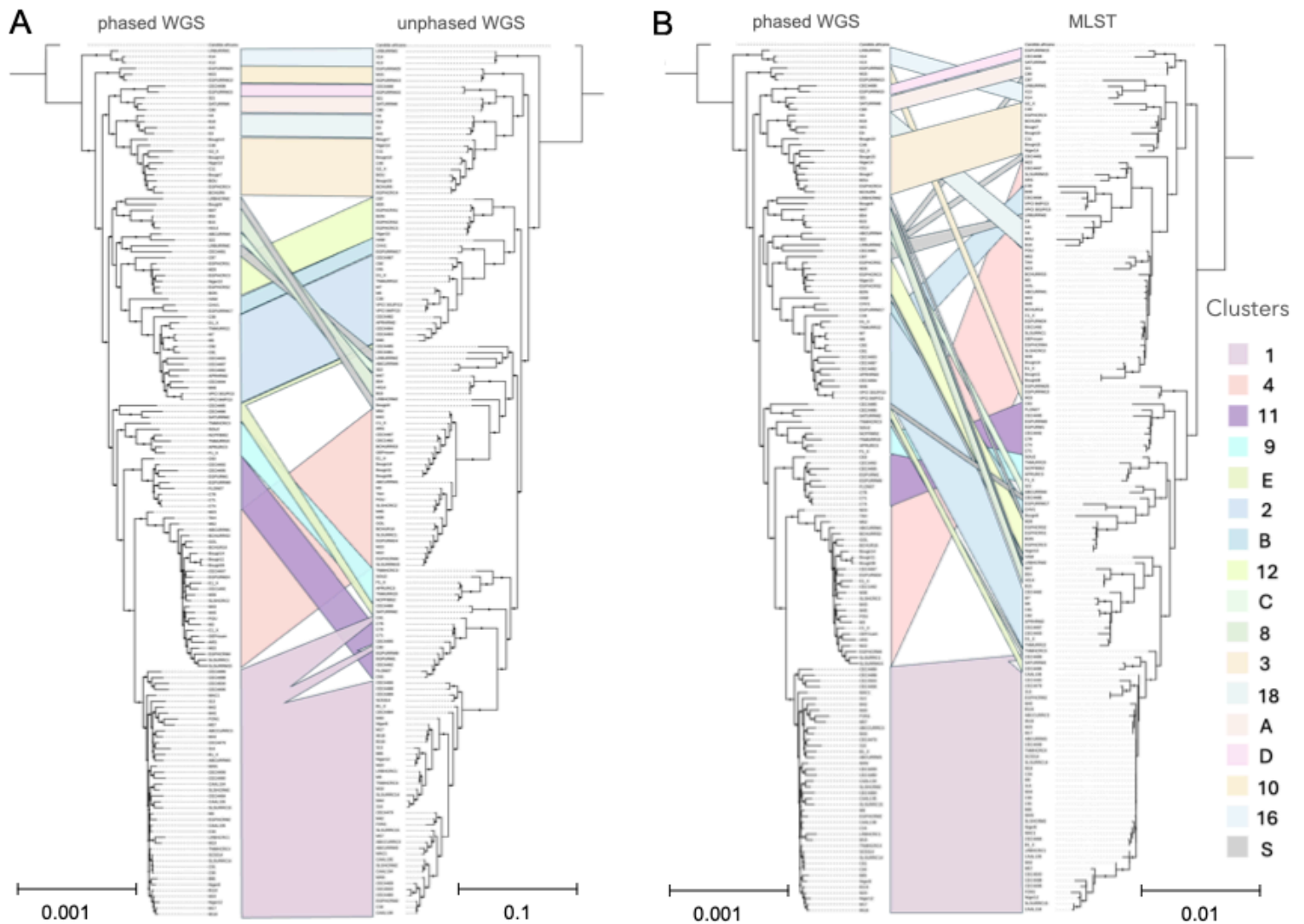


Figure 2

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 colour bands in both cases indicate a mismatch in the order of isolates and were drawn to assist with a visual interpretation of the differences. The black dots on the branches indicate bootstrap support values $\geq 80\%$. The legend colors are arranged according to the phased WGS genetic clusters.

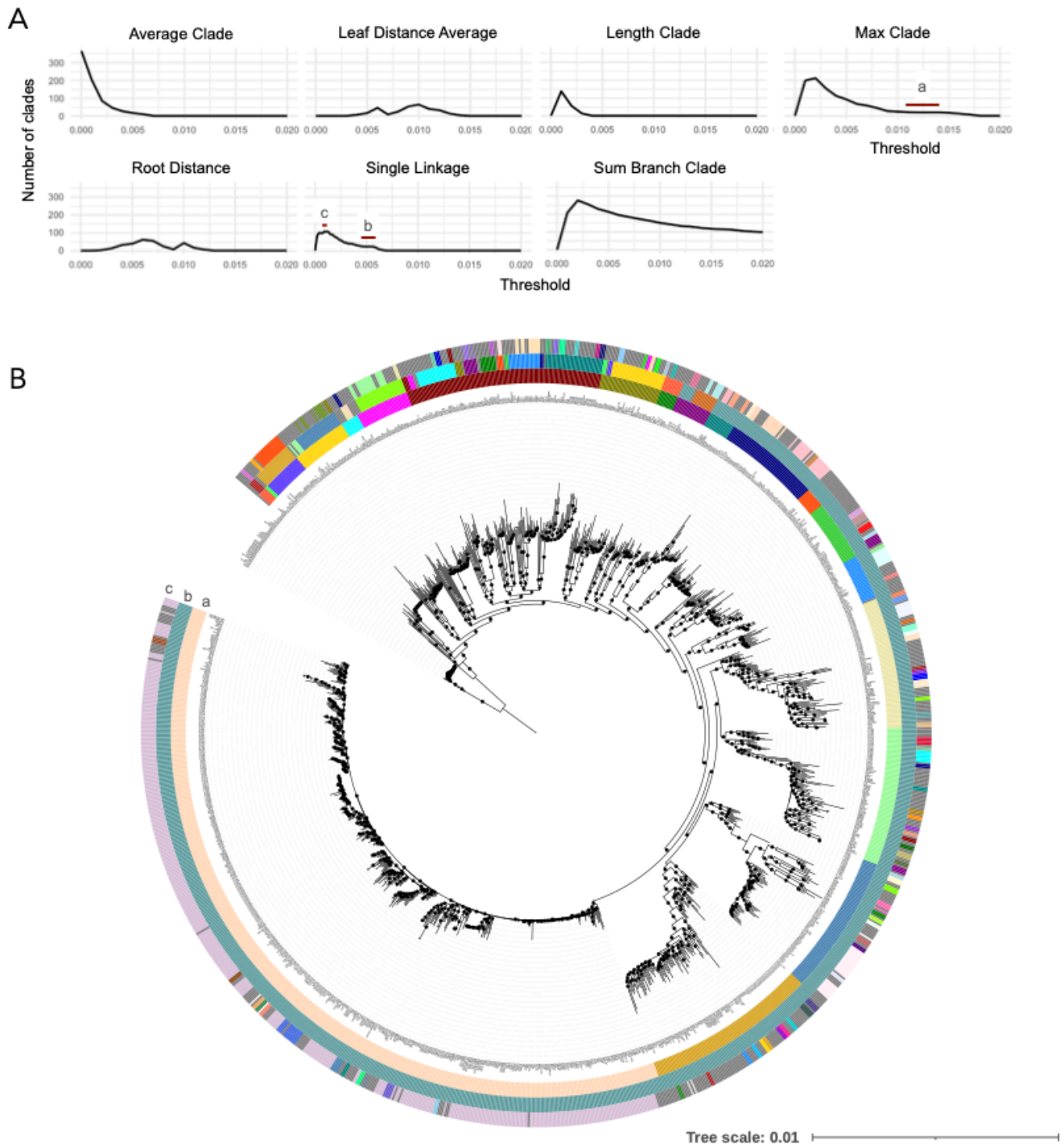


Figure 3

Clusters delineation using TreeCluster across multiple methods.

A) The number of predicted genetic clusters across a range of threshold values (0 to 1, in 0.001 increments) using seven TreeCluster methods. Regions with stable clusters predictions-defined as consistent cluster sizes across multiple consecutive thresholds, are highlighted with a dark brown bar

and labelled a–c. These include one max clade cluster (20 clusters, thresholds $t = 0.012 - 0.014$, a) and two single linkage clusters (24 clusters, b, and 109 clusters, c). B) Visualization of the stable clusters identified in (A) mapped onto the phylogenetic tree. Single linkage at a value of 0.009, which identified 109 clusters, was examined to visualize one of the high-cluster-number models and was not otherwise considered.

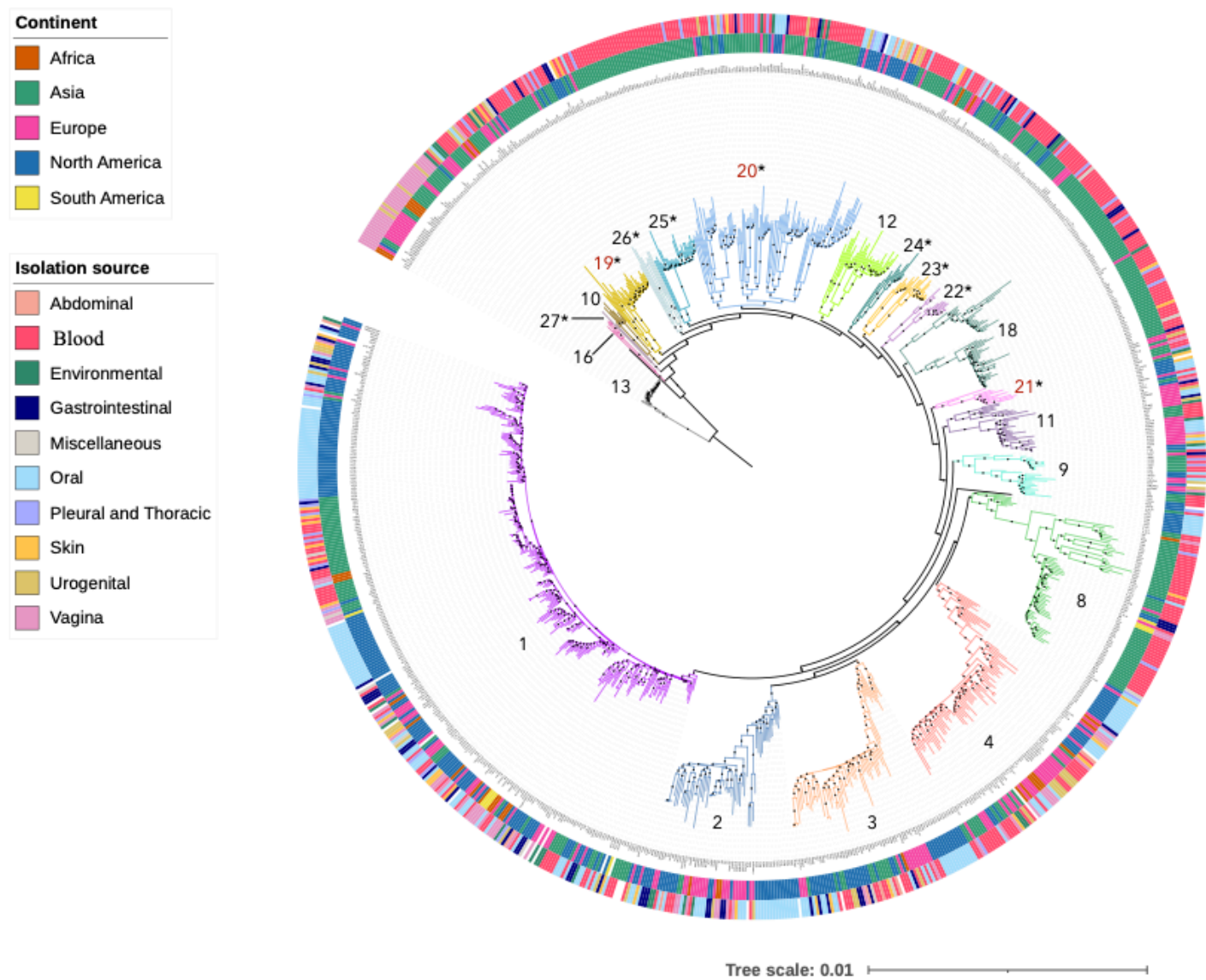


Figure 4

Maximum likelihood phylogeny of 938 isolates.

Branch colours correspond to cluster designations as defined in the study. The cluster numbers are denoted on the clusters, and asterisks are used to show new cluster labels. The red clusters are labels given to existing alphabetically designated clusters. Bootstrap support values $\geq 80\%$ are indicated by ribbons on the branches. Two concentric rings surround the phylogeny: the inner ring represents the

continent of origin for each isolate, while the outer ring denotes the anatomical source of isolation. This visualization highlights both the phylogenetic structure and the geographic and ecological diversity across clusters.

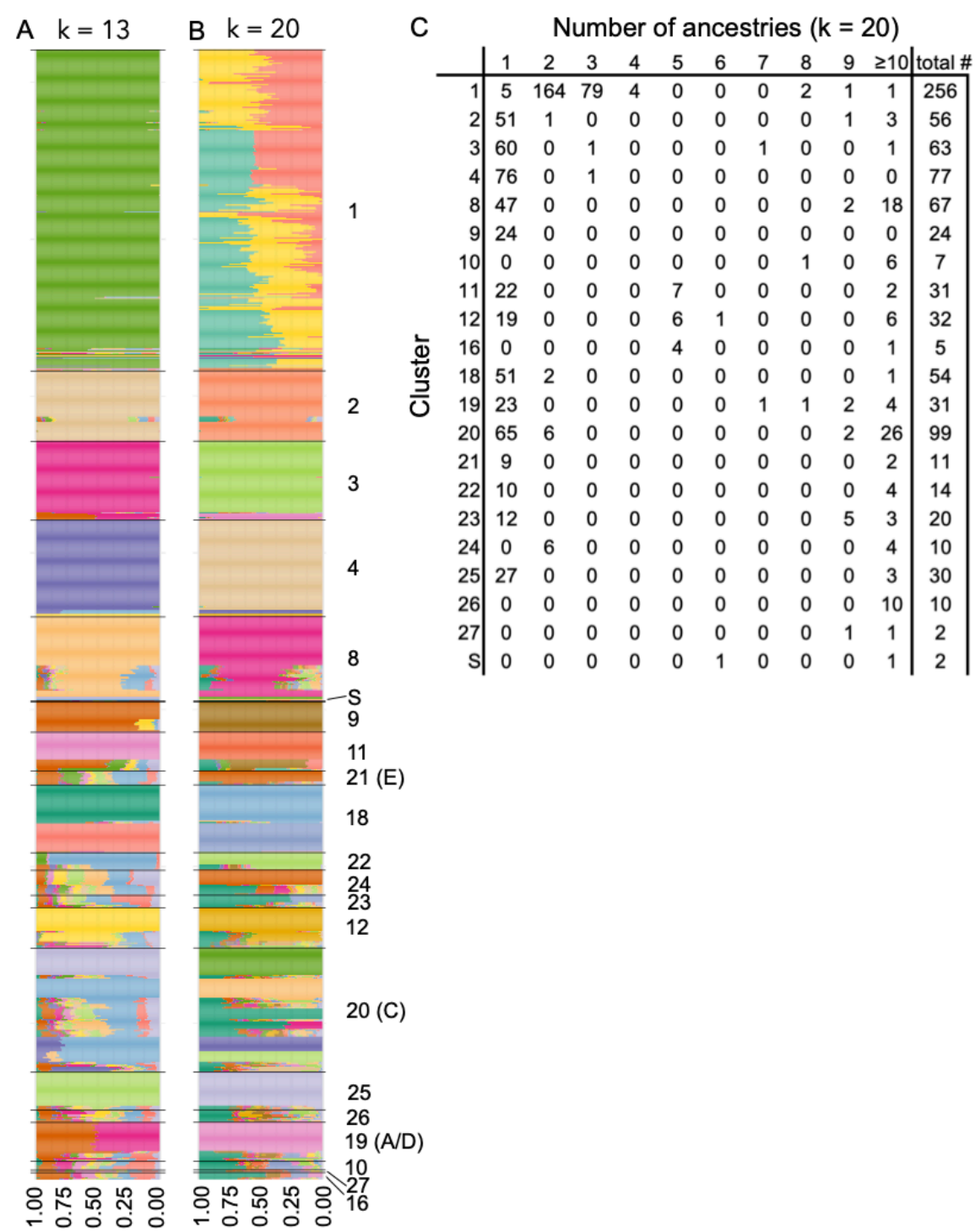
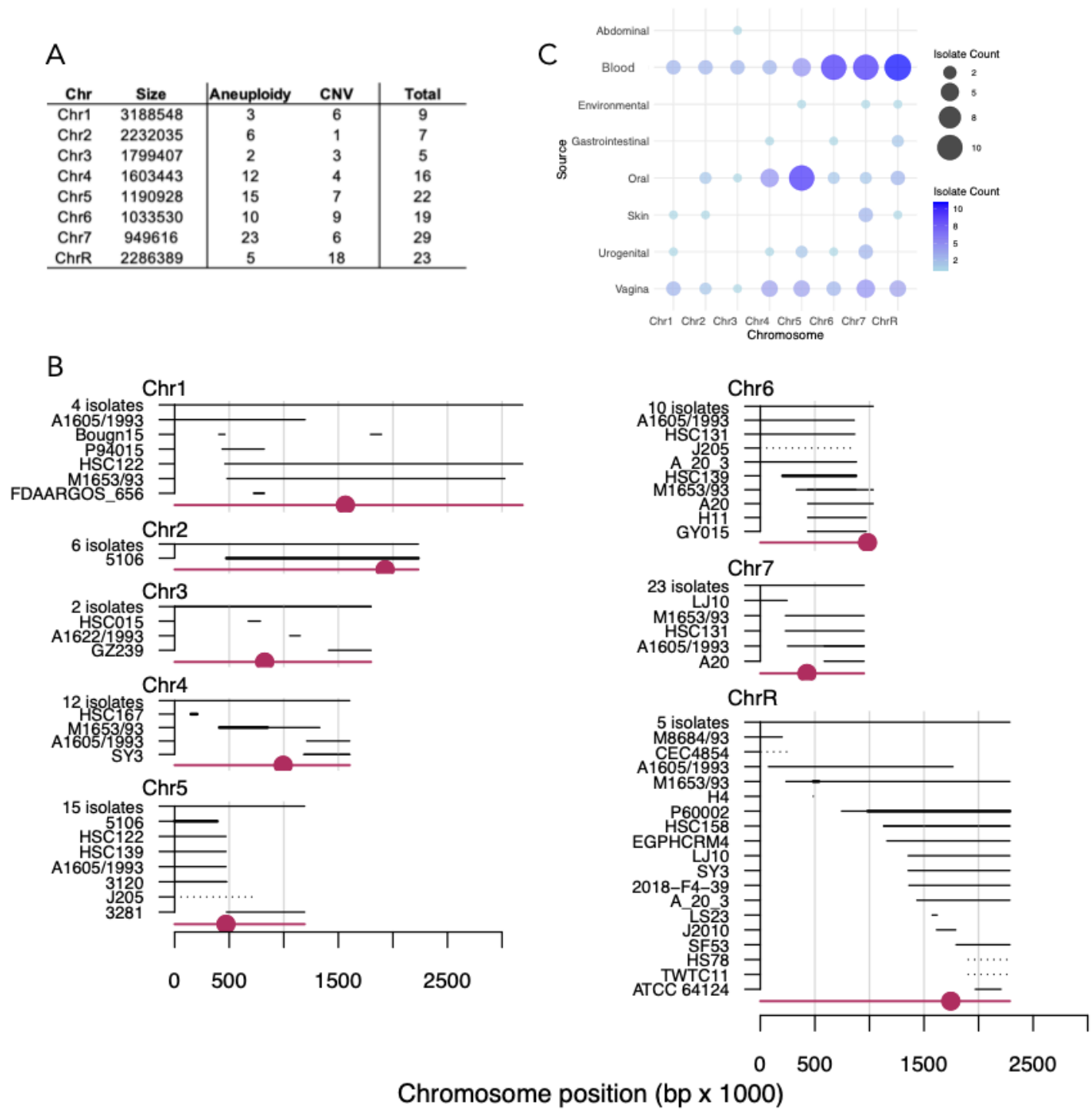


Figure 5

Population structure of the phylogenetically informative isolates inferred by NGSadmixture analysis.

NGSadmixture ancestry profiles of isolates at (A) $k = 13$ and (B) $k = 20$. Each horizontal bar represents an

individual isolate, ordered according to its position in the phylogenetic tree, and colored segments indicate the proportion of ancestry assigned to each inferred cluster. Horizontal black lines denote boundaries between proposed phylogenetic clades. (C) The table summarizes, for each clade, the number of distinct ancestry components detected among isolates at $k = 20$.



(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

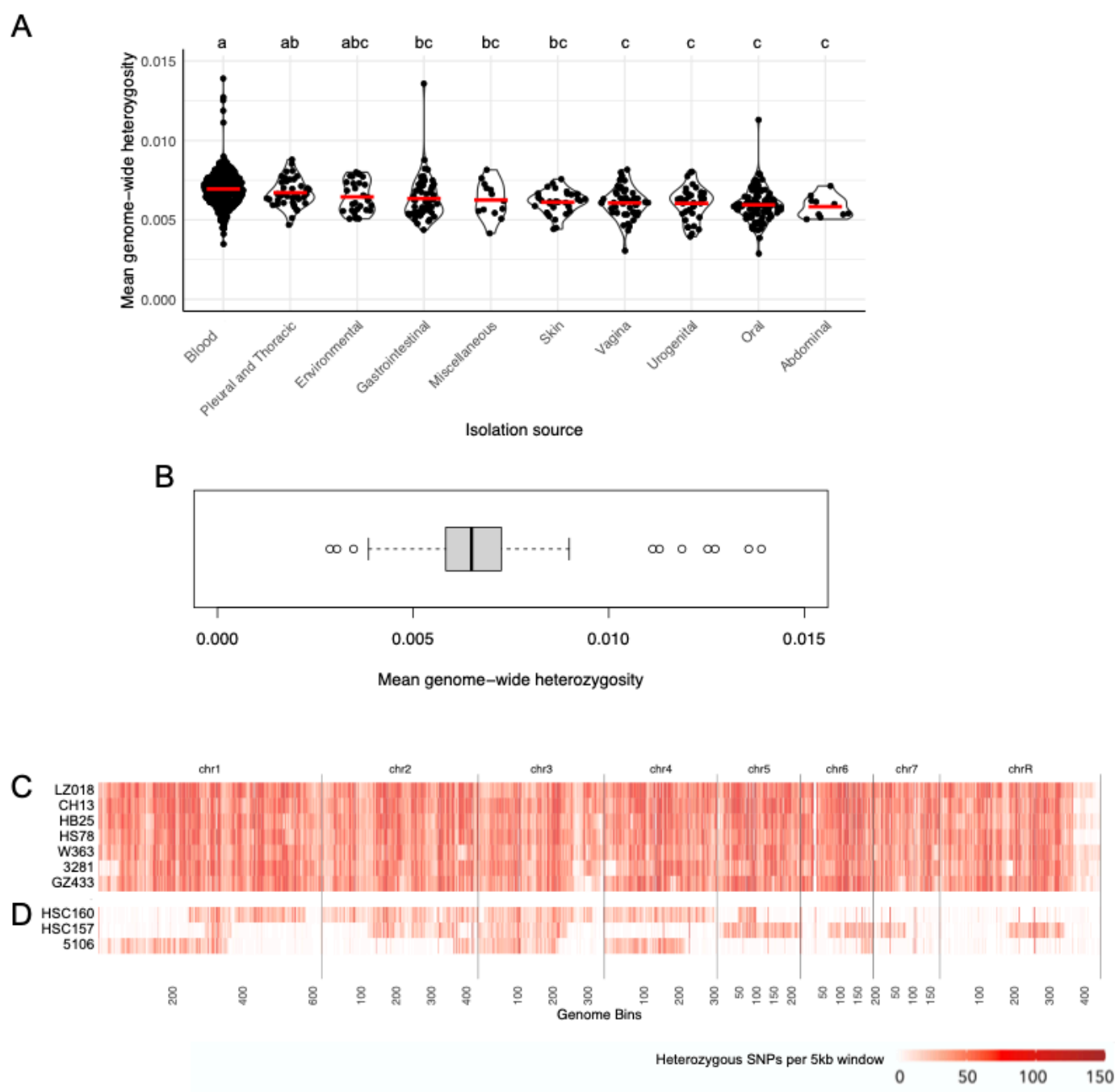


Figure 7

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) Outlier isolates with high average genome-wide heterozygosity. D) Outlier isolates with low heterozygosity, largely due to large LOH regions.

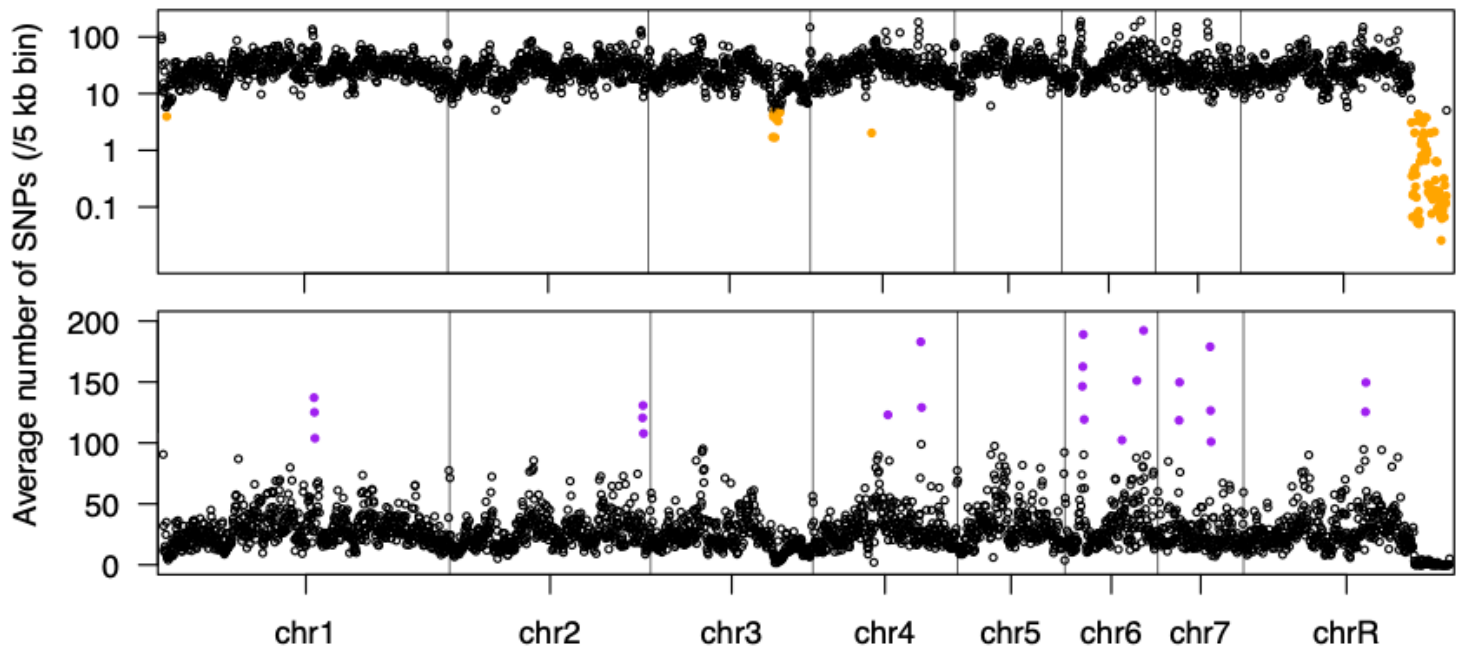


Figure 8

Genome-wide distribution of the average number of heterozygous SNPs across 745 *C. albicans* isolates.

Low-heterozygosity regions (defined as < 5 SNPs per window) are shown as orange circles (upper panel), while high-heterozygosity regions (defined as > 100 SNPs per window) are indicated by purple circles (lower panel).

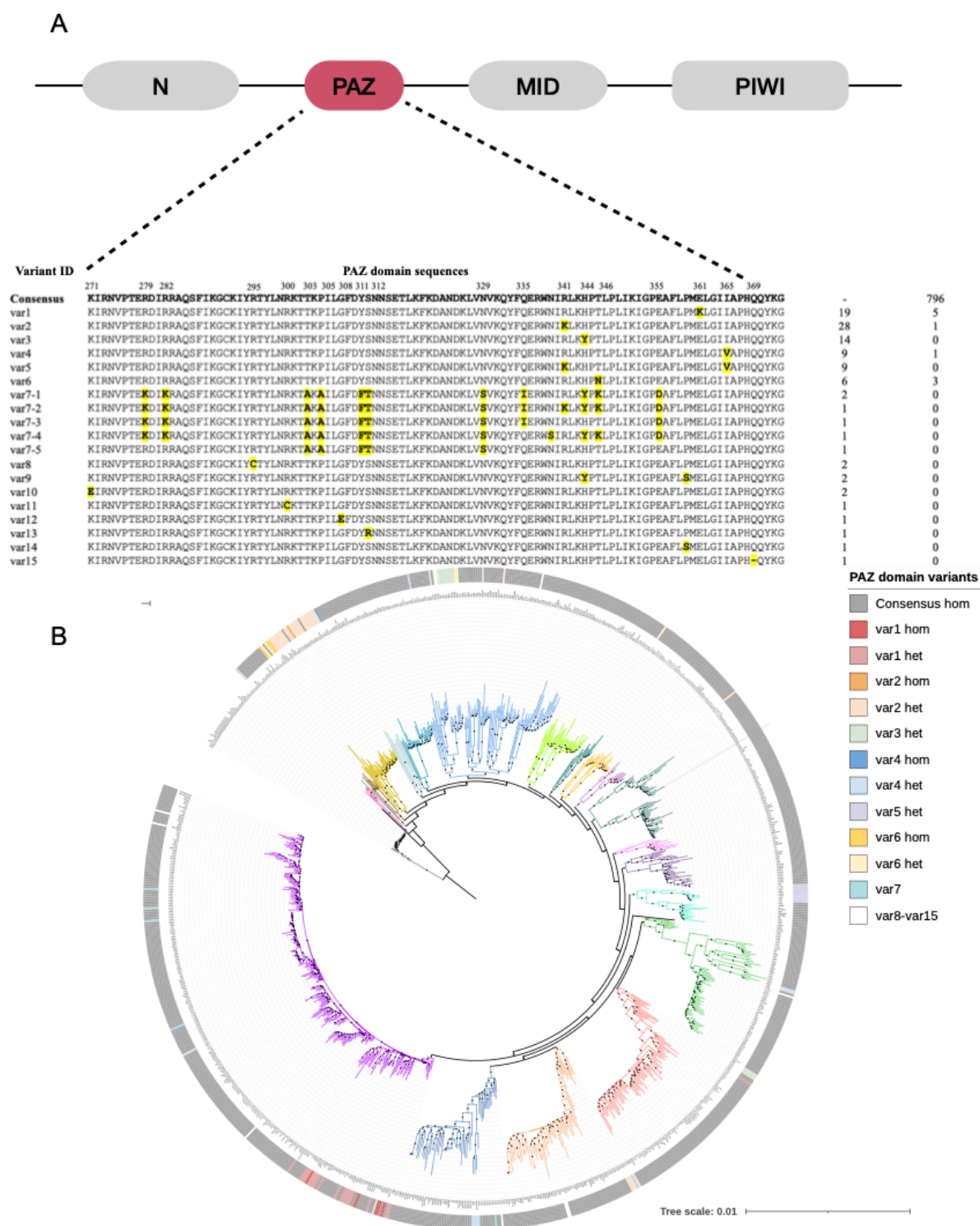


Figure 9

Candida albicans Ago1 PAZ domain sequence variants.

A) Schematic representation of *C. albicans* Ago1 protein structure and alignment of unique PAZ domain sequences (aa 271-373) identified from 908 *C. albicans* isolates with the consensus (known active) sequence in bold. Positions that differ from the consensus domain are highlighted in yellow, with the

positions indicated on the column. B) The distribution of the variant in the phylogeny. The branches of the phylogeny are colored by cluster.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TableS4.xlsx](#)
- [TableS2.xlsx](#)
- [TableS3.xlsx](#)
- [TableS1.xlsx](#)
- [SupplementaryFigures.pdf](#)