Parallel Expansion and Divergence of an Adhesin Family in Pathogenic Yeasts

Rachel A. Smoak*[1], Lindsey F. Snyder*[2], Jan S. Fassler^[3], Bin Z. He^[3]

[1]Civil and Environmental Engineering, [2]Interdisciplinary Graduate Program in Genetics, [3]Biology Department, University of Iowa, Iowa City, IA 52242

*These authors contributed equally

^Correspondence should be addressed to:

Bin Z. He, bin-he@uiowa.edu

Jan S. Fassler: jan-fassler@uiowa.edu

Abstract

Opportunistic yeast pathogens arose multiple times in the Saccharomycetes class, including the recently emerged, multidrug-resistant *Candida auris*. We show that homologs of a known yeast adhesin family in *Candida albicans*, the Hyr/Iff-like (Hil) family, are enriched in distinct clades of *Candida* species as a result of multiple, independent expansions. Following gene duplication, the tandem repeat-rich region in these proteins diverged extremely rapidly and generated large variations in length and β-aggregation potential, both of which were known to directly affect adhesion. The conserved N-terminal effector domain was predicted to adopt a β-helical fold followed by an α-crystallin domain, making it structurally similar to a group of unrelated bacterial adhesins. Evolutionary analyses of the effector domain in *C. auris* revealed relaxed selective constraint combined with signatures of positive selection, suggesting functional diversification after gene duplication. Lastly, we found the Hil family genes to be enriched at chromosomal ends, which likely contributed to their expansion via ectopic recombination and break-induced replication. Combined, these results suggest that the expansion and diversification of adhesin families generate variation in adhesion and virulence within and between species and are a key step toward the emergence of fungal pathogens.

Running title: Parallel evolution of adhesins in yeasts

Keywords: *Candida auris*, adhesin, gene duplication, convergent evolution, opportunistic yeast pathogen, natural selection

Introduction

Candida auris, a newly emerged multidrug-resistant yeast pathogen, is associated with a high mortality rate (up to 60% in a multi-continent meta-analysis (Lockhart et al. 2017)) and has caused multiple outbreaks across the world (CDC global *C. auris* cases count, February 15th, 2021). As a result, it became the first fungal pathogen to be designated by CDC as an urgent threat (CDC 2019). The emergence of *C. auris* as a pathogen is part of a bigger evolutionary puzzle: *Candida* is a polyphyletic genus that contains most of the human yeast pathogens. Phylogenetically, species like *C. albicans*, *C. auris* and *C. glabrata* belong to distinct clades with close relatives that either don't or only rarely infect humans (Fig. 1A). This strongly suggests that the ability to infect humans evolved multiple times in yeasts (Gabaldón et al. 2016). Because many of the newly emerged *Candida* pathogens are either resistant or can quickly evolve resistance to antifungal drugs (Lamoth et al. 2018; Srivastava et al. 2018), it is urgent to understand how yeast pathogenesis arose and what increases their survival in the host. We reason that any shared genetic changes among independently derived *Candida* pathogens will reveal key factors for host adaptation.

Gene duplication and the subsequent functional divergence is a major source for the evolution of novel phenotypes (Zhang 2003; Qian and Zhang 2014; Kuang et al. 2016; Eberlein et al. 2017). In a genome comparison of seven pathogenic *Candida* species and nine low-pathogenic potential relatives, three of the top six pathogen-enriched gene families encode GlycosylPhosphatidylInositol (GPI)-anchored cell wall proteins, namely Hyr/Iff-like, Als-like and Pga30-like (Butler et al. 2009). The first two encode known fungal adhesins (Bailey et al. 1996; Hoyer 2001; Luo et al. 2010). These glycosylated cell wall proteins play key roles in fungal attachment to host endo- and epithelial cells, mediate biofilm formation and iron acquisition, and are well-established virulence factors (HOYER et al. 2008; de Groot et al. 2013; Lipke 2018). It has been suggested that expansion of the cell wall protein repertoire, particularly adhesins, is a key step towards the evolution of yeast pathogens (Gabaldón et al. 2016). This is supported by a study showing that several adhesin families independently expanded in *C. glabrata* and close relatives (Gabaldón et al. 2013). Interestingly, studies of pathogenic *Escherichia coli* found that multiple strains independently acquired genes mediating intestinal adhesion, giving credence to the hypothesis from a different kingdom (Reid et al. 2000).

Despite the importance of adhesins in both the evolution and virulence of *Candida* pathogens, there is a lack of detailed phylogenetic study for their evolutionary history (Hoyer 2001; Linder and Gustafsson 2008; Gabaldón et al. 2013). Even less is known about their sequence divergence and the role of natural selection in their evolution (Xie et al. 2011). In the newly emerged *Candida auris*, individual adhesins have been characterized but there is little information about their evolutionary relationship with homologs in other *Candida* species and how their sequences diverged (Kean et al. 2018; Singh et al. 2019; Muñoz et al. 2021). In this study we characterized the detailed evolutionary history of a yeast adhesin family and used *C. auris* as a focal group to determine how adhesin sequences diverged under various natural selection forces. To choose a candidate adhesin family in *C. auris*, we compared it with the well-studied *C. albicans*, which belongs to the same CUG-Ser1 clade. Of the known adhesins in *C. albicans*, *C. auris* lacks the Hwp family and has only three Als or Als-like proteins compared with eight Als proteins in *C. albicans* (Muñoz et al. 2018). By contrast, *C. auris* has eight genes with a Hyphal_reg_CWP (PF11765) domain found in the Hyr/Iff family in *C. albicans* (Muñoz et al. 2021). This family was one of the most highly enriched in pathogenic *Candida* species relative to non-pathogenic ones (Butler et al. 2009). Transcriptomic studies identified two *C. auris* Hyr/Iff-like (Hil) genes as being upregulated during biofilm formation and under antifungal treatment (Kean et al. 2018). Interestingly, isolates from the less virulent *C. auris* Clade II lack five of the eight Hil genes (Muñoz et al. 2021). It is currently not known whether the *C. auris HIL* genes encode adhesins, how they relate to the *C. albicans* Hyr/Iff family genes and how their sequences diverged after duplication.

We show that the Hil family independently expanded multiple times, including in *C. auris* and *C. albicans*. Using *C. auris* as a focal species, we show in detail how sequence features and predicted structures of the effector domain offer support for the hypothesis that its Hil family members encode adhesins, while rates of nonsynonymous-to-synonymous substitutions reveal varying strengths of selective constraint and positive selection acting on the effector domain during the expansion of the family. The observed pattern of rapid divergence in the repeat-rich central domain was found to be general across the entire family and led to large variations in length and β-aggregation potential both between and within species, likely contributing to phenotypic diversity in adhesion and virulence.

Results

Phylogenetic distribution of the Hyr/Iff-like (Hil) family and its potential to encode adhesin

The Hyr/Iff family was first identified and characterized in *Candida albicans* (Bailey et al. 1996; Richard and Plaine 2007). The family is defined by its N-terminal Hyphally regulated Cell Wall Protein domain (Hyphal_reg_CWP, PF11765), followed by a highly variable central domain rich in tandem repeats (Boisramé et al. 2011). Because the effector domain is more conserved than the repeat region and plays a prominent role in mediating adhesion in known yeast adhesins (Willaert 2018), here we define the Hyr/Iff-like (Hil) family as the group of evolutionarily related proteins sharing the Hyphal_reg_CWP domain, different from a previous definition based on sequence similarity in either the Hyphal_reg_CWP domain or the repeat region (Butler et al. 2009).

To determine the phylogenetic distribution of the Hil family and its association with the pathogenic potential of species, we performed BLASTP searches using the Hyphal_reg_CWP domain from three distantly related Hil homologs as queries (from *C. auris*, *C. albicans* and *C. glabrata*). We scrutinized the database hits and searched additional assemblies to ensure that their sequences are complete and accurate given the available genome assemblies (Text S1). Using the criteria of E-value<10^{-5} and query coverage>50%, we identified a total of 215 proteins containing the Hyphal_reg_CWP domain from 32 species (Fig. 1A, Table S1). No credible hits were identified outside the budding yeast subphylum even after a lower E-value cutoff of 10^{-3} was tested, suggesting that this family is specific to this group (Materials and Methods). Species with eight or more Hil family genes fell largely within the Multi-Drug Resistant (MDR) and the Candida/Lodderomyces (CaLo) clades, which include *C. auris* and *C. albicans*, respectively. Only three such species were found outside of the two clades: *C. glabrata*, *M. bicuspidata* and *K. africana*. *C. glabrata* is a major opportunistic pathogen that is more closely related to *S. cerevisiae* than to most other *Candida* species (Dujon et al. 2004; Butler et al. 2009; Gabaldón et al. 2013). *M. bicuspidata* is part of the CUG-Ser1 clade. While not a pathogen in humans, it is a parasite of freshwater animals (Hall et al. 2010; Jiang et al. 2022). *K. africana* is not closely related to any known yeast pathogen and its ecology is poorly understood (Gordon et al. 2011).

We then asked how many of the Hil family genes in each species are likely to encode yeast adhesins. To get an initial estimate, we combined a Machine Learning tool for predicting fungal adhesins (Chaudhuri et al. 2011) with predictions for the N-terminal signal peptide and C-terminal GPI-anchor sequence, two features shared by the majority of known fungal adhesins (Lipke 2018). Half of all Hil homologs passed all three tests (Fig. 1A). Notably, *M. bicuspidata* has the largest Hil family among all species, but none of its 29 Hil genes passed all tests. We found most of the identified hits in this species were short relative to the rest of the family (Fig. S1), and 10 of the 29 hits were annotated as being incomplete in the RefSeq database. Further analyses with a better assembled genome and functional studies are needed to determine if the Hil family in this species has unique properties and functions.

Independent expansion of the Hil family in multiple pathogenic *Candida* lineages

Pathogenic yeast species have on average a larger Hil family and also more of its members were predicted to encode adhesins than in low pathogenic-potential species (Fig. 1B, t-test with unequal variance and Mann-Whitney U test both yielded *P* < 0.005, one-sided test). This naive comparison doesn't account for phylogenetic relatedness between species and could result in a false positive association (Levy et al. 2017; Bradley et al. 2018). To address this, we performed phylogenetic logistic regression, which uses the known phylogeny to specify the residual correlation structure among species with shared ancestry (Ives and Garland 2010). We tested for associations between the pathogen status with either the total number of Hil homologs or the number of putative adhesins in each species. Both tests were significant (*P* = 0.005 and 0.007, respectively). Together, these results strongly support an enrichment of the Hil family and the putative adhesins therein among the pathogenic yeast species.

Some adhesin families have undergone independent expansions even among closely related species (Gabaldón et al. 2013). This would result in overestimation of the phylogenetic signal in the above analysis. To further characterize the evolutionary history of the Hil family, including among closely related *Candida* lineages, we reconstructed a species tree-aware maximum likelihood phylogeny for the Hil family based on the Hyphal_reg_CWP domain alignment (Fig. 1C, Fig. S2). We found that homologs from the MDR clade and the Candida / Lodderomyces (CaLo) clade separated into two groups, suggesting that the duplications of the Hil family genes in the two clades occurred independently. To better illustrate the history of gene duplications in the Hil family, we reconciled the gene tree with the species tree and mapped the number of duplications onto the species phylogeny (Materials and Methods). The result showed that the Hil family has independently expanded multiple times, not only between clades but also among closely related species within a clade, such as in *C. albicans* and *C. tropicalis* (Fig. 1D).

Sequence features of the *C. auris* Hil family support their adhesin status

Experiments have demonstrated that Hil family members function as adhesin in *C. albicans* and more recently for one member in *C. glabrata* (Bailey et al. 1996; Boisramé et al. 2011; Reithofer et al. 2021; Rosiana et al. 2021). To further evaluate the adhesin function of Hil family proteins, we focused on *C. auris*, in which Hil family members were implicated in biofilm formation and response to antifungal treatments, but still remain poorly characterized (Kean et al. 2018). We named the eight *C. auris* Hil family proteins Hil1-Hil8 ordered by their length (Table S2). This differs from the literature, which referred to them by their most closely related Hyr/Iff genes in *C. albicans* (Kean et al. 2018; Jenull et al. 2021; Muñoz et al. 2021). The renaming was to avoid the incorrect implication of one-to-one orthology between the two species (Fig. 1C).

To further assess the adhesin potential for the *C. auris* Hil family, we compared their domain architecture and sequence features to those typical of known yeast adhesins, including a signal peptide, an effector domain, a Ser/Thr-rich and highly glycosylated central domain with tandem repeats and β-aggregation prone sequences and a GPI-anchor signal (Fig. 2A) (de Groot et al. 2013; Lipke 2018). All eight *C. auris* Hil proteins followed this domain architecture (Fig. 2B). Hil1-4 were additionally characterized by an array of regularly spaced β-aggregation prone sequences (red ticks below the protein, Fig. 2B). All eight proteins also had elevated Ser/Thr frequencies in their central domain and were predicted to be heavily O-glycosylated (Fig. 2C). Predicted N-glycosylation was rare except in Hil5 and Hil6 (Fig. 2C). The overall Ser/Thr frequencies in the Hil family proteins were significantly elevated compared with the rest of the proteome (Fig. S3). All eight members were predicted to be fungal adhesins by FungalRV, a support vector machine-based classifier that showed high sensitivity and specificity in eight pathogenic fungi based on sequence features (Chaudhuri et al. 2011).

Hyphal_reg_CWP domain in the Hil family is predicted to adopt a β-helical fold similar to unrelated bacterial adhesin binding domains

Crystal structures of the effector domain in several yeast adhesin families, including Als, Epa and Flo, revealed carbohydrate or peptide binding activities supporting the proteins' adhesin functions (Willaert 2018). The structure of the Hyphal_reg_CWP domain in the Hil family in this study has not yet been experimentally determined. However, crystal structures for the effector domains of two Adhesin-like Wall Proteins (Awp1 and Awp3b) in *C. glabrata*, which are distantly related to those in the Hil family, were recently reported, and the predicted structure of one of *C. glabrata*'s Hil family members (Awp2) was found to be highly similar to the two solved structures (Reithofer et al. 2021). We used AlphaFold2 (Jumper et al. 2021) to predict the structures of the effector domain for two *C. auris* Hil proteins, Hil1 and Hil7 (Fig. 3A, B). Both resemble the *C. glabrata* Awp1 effector domain (Fig. 3C), consisting of a right-handed β-helix at the N-terminus followed by an α-crystallin fold. There are three β-strands in each of the 9 rungs in the β-helix, stacked into three parallel β-sheets (Fig. 3D). The α-crystallin domain consists of seven β-strands forming two antiparallel β-sheets, adopting an immunoglobulin-like β-sandwich fold (Fig. 3E) (Koteiche and Mchaourab 1999; Stamler et al. 2005).

The β-strand-rich structure is typical of effector domains in known yeast adhesins, but the β-helix fold at the N-terminus is uncommon (Willaert 2018). Proteins with a β-helix domain often have carbohydrate-binding capabilities and act as enzymes, e.g., hydrolase and pectate lyase (SCOP ID: 3001746). To gain further insight into Hyphal_reg_CWP domain's function, we searched the PDB50 database for structures similar to what was predicted for *C. auris* Hil1 using DALI (Holm 2022). We identified a number of bacterial adhesins with a highly similar β-helix fold but no α-crystallin domain (Table S3), e.g., Hmw1 from *H. influenzae* (PDB: 2ODL), Tāpirins from *C. hydrothermalis* (PDB: 6N2C), TibA from enterotoxigenic *E. coli* (PDB: 4Q1Q) and SRRP from *L. reuteri* (PDB: 5NY0). For comparison, the binding region of the Serine Rich Repeat Protein 100-23 ($SRRP_{100-23}$) from *L. reuteri* was shown in Fig. 3F (Sequeira et al. 2018). Together, these results strongly suggest that the Hyphal_reg_CWP domain in the *C. auris* Hil family genes mediates adhesion. Additionally, the low sequence identity (12-15%) between the yeast Hyphal_reg_CWP domain and the bacterial adhesins' binding regions further suggests the two groups have convergently evolved a similar structure to achieve adhesion functions.

Rapid divergence of the repeat-rich central domain in Hil family proteins in *C. auris*

While the overall domain architecture is well conserved, the eight Hil family proteins in *C. auris* differ significantly in length and sequence of their central domains (Fig. 2B). While not involved in ligand binding, central domains in yeast adhesins are known to play a critical role in mediating adhesion: the length and stiffness of the central domain are essential for elevating and exposing the effector domain (Frieman et al. 2002; Boisramé et al. 2011); and the tandem repeats and β-aggregation sequences within them directly contribute to adhesion by mediating homophilic binding and amyloid formation (Rauceo et al. 2006; Otoo et al. 2008; Frank et al. 2010; Wilkins et al. 2018). Thus, divergence in the central domain of the Hil family has the potential to lead to phenotypic diversity, as shown in *S. cerevisiae* (Verstrepen et al. 2004; Verstrepen et al. 2005).

To determine how the central domain sequences evolved in the *C. auris* Hil family, we used dot plots both to reveal the tandem repeat structure within each protein and to examine the similarity among the paralogs. A "dot" on the x-y plot indicates that the corresponding segments (window size = 50 a.a.) from the two proteins on the x- and y-axes share similarity, with the gray scale being proportional to the degree of similarity (Brodie et al. 2004). We found that *C. auris* Hil1, 2, 3 and 4 share a ~44 aa repeat unit, whose copy number varies between 15 and 46, driving differences in their protein lengths (Fig. 4A). These repeats have conserved periodicity as well as sequence (Fig. 4B, Fig. S4). There are two interesting features of this 44 aa repeat unit: a) it contains a heptapeptide "GVVIVTT" that is predicted to be strongly β-aggregation prone, which explains the large number of regularly spaced β-aggregation motifs in Hil1-Hil4 (Fig. 2B); b) it is predicted to form three β-strands in the same orientation (Fig. 4B), raising an interesting question of whether the tandem repeats may adopt a β-structure similar to that of the effector domain. Hil7 and Hil8 encode the same repeat unit but have only one copy (Fig. 4A, red boxes). By contrast, Hil5 and Hil6 encode very different low complexity repeats with a unit length of ~5 aa. Their copy numbers range between 15 to 49 (Fig. 4C, D) have relatively low Ser/Thr frequencies (Fig. 2C). Another consequence of encoding only one or zero copies of the 44 aa repeat unit found in Hil1-Hil4 is that Hil5-Hil8 are predicted to have 2-4 β-aggregation prone sequences in contrast to 21-50 in Hil1-Hil4. For comparison, characterized yeast adhesins contain 1-3 such sequences at a cutoff of >30% β-aggregation potential predicted by TANGO (Fernandez-Escamilla et al. 2004; Ramsook et al. 2010; Lipke 2018). The variable lengths, Ser/Thr frequencies and distribution of β-aggregation sequences, all resulting from the evolution of the tandem repeats, suggest the intriguing possibility that the 8 different Hil proteins in *C. auris* are non-redundant, playing distinct roles in cell adhesion and other cell-wall related phenotypes.

Because tandem repeats are prone to recombination-mediated expansions and contractions, we asked if there are variable numbers of tandem repeats (VNTR) among strains in *C. auris*, which could generate diversity in cell adhesive properties as shown in *S. cerevisiae* (Verstrepen et al. 2005). To answer this question, we identified homologs of Hil1-Hil4 in nine *C. auris* strains from three geographically-stratified clades (Muñoz et al. 2018; Muñoz et al. 2021). The genomes of these strains were *de novo* assembled using long-read technologies (Table S4), which allowed us to confidently assess copy number variations within tandem repeats. We identified a total of eight indel polymorphisms in Hil1-Hil4 (Table S5, example alignments in Fig. S5). Except for one 16 aa deletion that is in a single Clade III strain, all seven other indels span one or multiples of the repeat unit and affect all strains within a clade. This is consistent with them being driven by recombination between repeats. The agreement within clades additionally show the indels are not due to sequencing / assembly artifacts, which are not expected to follow the clade labels. As previously reported, Clade II strains lack five of the eight Hil family proteins, including Hil1-4 (Muñoz et al. 2021). Our phylogenetic analysis further showed that this was due to gene losses within Clade II (Fig. S6). The potential relationship between the Hil family size and the virulence profiles of Clade II strains is discussed later.

Natural selection on the effector domain during the Hil family expansion in *C. auris*

Gene duplication provides raw materials for natural selection and is often followed by a period of relaxed functional constraints on one or both copies, allowing for sub- or neo-functionalization (Zhang 2003; Innan and Kondrashov 2010). Positive selection can be involved in this process, which can lead to a ratio of nonsynonymous to synonymous substitution rates dN/dS > 1 (Yang 1998). Here we ask if the Hyphal_reg_CWP domain in *C. auris* Hil1-Hil8 experienced relaxed selective constraints and/or positive selection following gene duplications, the latter of which would suggest functional diversification. We chose to focus on the Hyphal_reg_CWP domain because of its functional importance and because the high-quality alignment in this domain allowed us to make confident evolutionary inferences (Fig. S7).

Because gene conversion between paralogs can cause distinct genealogical histories for different parts of the alignment and mislead evolutionary inferences (Casola and Hahn 2009), we first identified putatively non-recombining partitions using GARD (Kosakovsky Pond et al. 2006) (Fig. S8), and chose two partitions, P1-414 and P697-981, for maximum-likelihood based analyses using PAML (Yang 2007) (Fig. 5A).

We first tested if a subset of the sites evolved under positive selection consistently on *all* branches. We found moderate evidence supporting the hypothesis for the P697-981 partition, where the M8 vs. M7 and M8 vs. M8a tests were significant at a 0.01 level, but the conservative test M2a vs M1a was not (Table S6). All three tests were insignificant for the P1-414 partition. Next, we tested for elevated dN/dS on selected branches of the tree, sign of relaxed selective constraints or positive selection. We first estimated the dN/dS for each branch using a free-ratio model and designated those with dN/dS greater than 10 as the "foreground" (Fig. 5B, C, "FG"). We found strong evidence for the FG branches to have a higher dN/dS than the remainder of the tree (log-likelihood ratio test $P < 0.01$, Fig. 5D). There is no evidence, however, for the dN/dS across the entire domain on the FG branches to be greater than 1 (Fig. 5D, *a*, row 2). We then tested the more realistic scenario, where a subset of the sites on the FG branches were subject to positive selection. Using the branch-site test 2 as defined in (Zhang et al. 2005), we found evidence for positive selection on a subset of the sites on the FG branches for both partitions (log-likelihood ratio test $P < 0.01$), and identified residues in both as candidate targets of positive selection with a posterior probability greater than 0.99 (Fig. 5D). We conclude that there is strong evidence for relaxed selective constraint on the Hyphal_reg_CWP domain on some branches following gene duplications; there is also evidence for positive selection acting on a subset of the sites on those branches. However, as the free-ratio model estimates were noisy and the Empirical Bayes method used to identify the residues under selection lacks power (Zhang et al. 2005) and can produce false positives (Nozawa et al. 2009), the specific branches and residues implicated must be interpreted with caution.

The yeast Hil family has adhesin-like domain architecture with rapidly diverging central domain sequences

We next examined the entire yeast Hil family to reveal the broader patterns of their evolution. We found that the Hil family in general has elevated Ser/Thr content compared with the rest of the proteome (Fig. S9). Moreover, the majority encode tandem repeats in the central domain (Fig. 6A) and contain predicted β-aggregation prone sequences (Fig. 6B). Together, these features further suggest that most yeast Hil family members encode fungal adhesins. While these key features typical of yeast adhesins are conserved, the yeast Hil family exhibits extreme variation in protein length, tandem-repeat content as well as in β-aggregation potential (Fig. 6A, B, S10), extending the pattern seen in *C. auris* (Fig. 2). The length of the protein outside of the

Hyphal_reg_CWP domain has a mean ± standard deviation of 822.4±785.8 aa and a median of 608.5 aa. This large variation in protein length is almost entirely driven by the tandem repeats (Fig. 6C, linear regression slope = 1.0, r^2 = 0.83). A subset of the Hil proteins (vertical bar in Fig. 6A, B) stand out in that they are both longer than the rest of the family (1745 vs 770 aa, median protein length) and have an unusually large number of β-aggregation prone motifs (25 vs 6, median number of TANGO hits per protein). The motifs in this group of proteins are regularly spaced as a result of being part of the tandem repeat unit (median absolute deviation, or MAD, of distances between adjacent TANGO hits less than 5 aa, Fig. 6D). The motif "GVVIVTT" and its variants account for 61% of all hits in this subset and are not found in significant number in the rest of the family. Together, these observations combined with previous experimental studies showing a direct impact of adhesin length and β-aggregation potential on their function (Verstrepen et al. 2005; Lipke et al. 2012) lead us to propose that the rapid divergence of the Hil family following the parallel expansion led to functional diversification in adhesion in pathogenic yeasts and may have contributed to their enhanced virulence.

The yeast Hil family genes are preferentially located near chromosome ends

Several well-characterized yeast adhesin families, including the Flo family in *S. cerevisiae* and the Epa family in *C. glabrata*, are enriched in the subtelomeres (Teunissen and Steensma 1995; De Las Peñas et al. 2003; Xu et al. 2020; Xu et al. 2021). This region is associated with high rates of SNPs, indels and copy number variations, and can undergo ectopic recombination that enables the spread of genes between chromosome ends or their losses (Mefford and Trask 2002; Anderson et al. 2015). To determine if the Hil family is also enriched in the subtelomeric region, we compared their chromosomal locations with the background gene density distribution (Fig. 7A) in species with a chromosomal level assembly (Table S7). To account for the shared evolutionary history, we selected one species per closely related group such that the Hil family homologs in these species were mostly derived through independent duplications based on our gene tree (Fig. S2). The result showed that the Hil family genes are indeed enriched at chromosomal ends (Fig. 7B). A goodness-of-fit test confirmed that the difference between the chromosomal locations of the Hil family and the genome background is highly significant ($P = 1.3 \times 10^{-12}$). As ectopic recombination between subtelomeres has been suggested to underlie the spread of gene families (Anderson et al. 2015), we hypothesize that the enrichment of the Hil family towards the chromosome ends is both a cause and consequence of its parallel expansion in different *Candida* lineages.

Discussion

The repeated emergence of human pathogens in the Saccharomycetes class poses serious health threats, as many emerging pathogenic species are multi-drug resistant or quickly gain resistance (Lamoth et al. 2018; Srivastava et al. 2018). This raises an evolutionary question: are there shared genomic changes in independently derived *Candida* pathogens, which could be key factors in host adaptation? Yeast adhesin families were among the most enriched gene families in pathogenic lineages relative to the low pathogenic potential relatives (Butler et al. 2009). It has been proposed that expansion of adhesin families could be a key step in the emergence of novel yeast pathogens (Gabaldón et al. 2016). However, detailed phylogenetic studies supporting this hypothesis are rare (Gabaldón et al. 2013), and far less is known about how their sequences diverge and what selective forces are involved during the expansions (Xie et al. 2011; Muñoz et al. 2021). In this study, we found that the Hyr/Iff-like (Hil) family, defined by the conserved Hyphal_reg_CWP domain, is significantly enriched among distantly related pathogenic clades (Fig. 1A, B). This resulted from independent expansion of the family in these clades, including among closely related species (Fig. 1C, D). We also showed that the protein sequences diverged extremely rapidly after duplications, driven mostly by the evolution of the tandem repeats and resulting in large variations in protein length, Ser/Thr content and β-aggregation potential (Fig. 2B, C, Fig. 6). Our evolutionary analyses revealed evidence of relaxed selective constraint and a potential role of positive selection acting on the Hyphal_reg_CWP domain during the family's expansion in *C. auris* (Fig. 5). We also found the Hil family to be strongly enriched near chromosomal ends (Fig. 7). Overall, our results support the hypothesis that expansion and diversification of adhesin families is a key step towards the emergence of yeast pathogens.

Genome assembly quality limits gene family evolution studies

Like any study of multi-gene family evolution, our work relies on and is limited by the quality of the genome assemblies. Two additional challenges in our study are due to the fact that Hil family genes are rich in tandem repeats (Fig. 2B, 6A), and many are located near chromosome ends (Fig. 7B), both of which pose problems for genome assemblies. For example, we found significant disagreement in length for 8 of the 16 Hil proteins in *C. tropicalis* between a long-read assembly and the RefSeq assembly, consistent with a recent study (Oh et al. 2020) (Table S8); in *C. glabrata*, we identified 13 Hil family genes in a long-read assembly (GCA_010111755.1) vs 3 in the RefSeq assembly (GCF_000002545.3); 12 of the 13 genes were in the subtelomeres (Xu et al. 2020). However, similar analyses in additional species didn't reveal these problems, suggesting that the issues were at least in part due to difficulties in some genomes (Text S1). Nonetheless, we acknowledge the possibility of missing homologs and inaccurate sequences, especially in the tandem-repeat region. We thus believe the expected improvements in genome assemblies due to advances in long-read sequencing technologies will be crucial for future studies of the adhesin gene family in yeasts. It is worth noting that our main conclusions about the parallel expansion of the Hil family and its rapid divergence patterns are robust with respect to isolated problems as described above. Also, the long-read technology-based and *de novo* assembled genomes for *C. auris* strains allowed us to confidently assess variation in the Hil family size and tandem repeat copy number between paralogs and among individual strains (Table S4). The accuracy of the tandem-repeat sequences in multiple strains in this species is supported by the conservation of repeat copy numbers within clades (Table S5).

Evidence for adhesin functions in the Hil family

A few members of the Hil family, e.g., Iff4 in *C. albicans* and Awp2 in *C. glabrata* were shown to mediate adhesiveness to polystyrene (Fu et al. 2008; Kempf et al. 2009; Reithofer et al. 2021). While further experimental studies are needed to establish the adhesin functions of other Hil family members, our work provides bioinformatic support for this hypothesis (Fig. 2, 6). The predicted β-helix fold of the Hyphal_reg_CWP domain (Fig. 3), while unusual among characterized yeast adhesins (Willaert 2018), is found in many virulence factors residing on the surface of bacteria or viruses as well as enzymes that degrade or modify polysaccharides (Table S3) (Kajava and Steven 2006). The elongated shape and rigid structure of the β-helix are consistent with the functional requirements of adhesins, including the need to protrude from the cell surface and the capacity for multiple binding sites along its length that facilitate adhesion. In a bacterial adhesin – the serine rich repeat protein (SRRP) from the Gram-positive bacterium, *L. reuteri* – a protruding, flexible loop in the β-helix was proposed to serve as a binding pocket for its ligand (Sequeira et al. 2018). Such a feature is not apparent in the predicted structure of the Hyphal_reg_CWP domain. Further studies are needed to elucidate the mechanism of action of this domain and its potential substrates.

The cross-kingdom similarity in adhesin effector domain structure is intriguing in several ways. First, it suggests convergent evolution in bacteria and yeasts. Second, it suggests that what is known about the structure-function relationship in bacteria can provide insight into the Hyphal_reg_CWP domain in yeast. Notably, the *Lr*SRRP shows a pH-dependent substrate specificity that is potentially adapted to distinct host niches (Sequeira et al. 2018). Finally, the similar structure and function of the bacterial and yeast adhesins could mediate cross-kingdom interactions in natural and host environments (Uppuluri et al. 2018).

However, not all Hil family homologs are likely to encode adhesins. Sequence features suggest some Hil family proteins may have non-adhesin functions. For example, 39 of 193 Hil proteins (homologs labeled as incomplete were excluded) have the requisite signal sequence (SP+), but lack a GPI anchor attachment site (gpi-, Fig. S1). One, Iff11 in *C. albicans*, was shown to be secreted, and a null mutant of it was found to be hypersensitive to cell wall-damaging agents and less virulent in a murine systemic infection model (Bates et al. 2007). Moreover, 75% of these "SP+, gpi-" proteins are shorter than 600 amino acids, in contrast to only 4% of the 117 proteins having both a signal peptide and a GPI anchor attachment site. Such short, secreted proteins with tandem repeat sequences identical or similar to those present in the cell-wall associated Hil protein counterparts may serve an important regulatory function by bundling with wall associated adhesins as previously suggested for similar subclass of proteins within the Als family (Oh et al. 2019). It is possible that the Hil family has evolved diverse functions broadly related to cell adhesion.

Ongoing diversification of the Hil family within species

In addition to the parallel expansion and the subsequent rapid sequence divergence in the Hil family between species, we and others also revealed population level variation in both the family size and sequences within *C. auris* (Fig. S5, S6, Table S5) (Muñoz et al. 2021). Notably, among the four geographically stratified clades, Clade II strains lost five of the eight Hil family members (Fig. S6). Besides missing members of the Hil family, Clade II strains also lack seven of the eight members of another GPI-anchor family that is specific to *C. auris* (Muñoz et al. 2021). These coincide with the finding that Clade II strains were mostly associated with ear infections (57/61 isolates according to (Kwon et al. 2019)) rather than hospital outbreaks, as reported for strains from the other clades, and that they were generally less resistant to antifungal drugs (Kwon et al. 2019; Welsh et al. 2019). This raises the question of whether the smaller adhesin repertoire in Clade II strains limits their adhesive capability and results in a different pathology. Similar expansion and contraction of adhesin families have been shown for the *C. glabrata* Hil family (AWP Cluster V) and Epa family (Marcet-Houben et al. 2022), suggesting that dynamic evolution of adhesin families in pathogenic yeasts could be a common pattern. Variation in the tandem repeat copy number in Hil1-Hil4 among *C. auris* strains is also intriguing (Fig. S5). Prior studies of the *S. cerevisiae* Flo proteins have shown that protein length directly impacts cellular adhesion phenotypes (Verstrepen et al. 2005) and thus population level variation in adhesin length could further contribute to phenotypic diversity. Lastly, scans for selective sweeps in *C. auris* identified Hil and Als family members as being among the top 5% of all genes, suggesting that adhesins are targets of natural selection in the recent evolutionary history of this newly emerged pathogen (Muñoz et al. 2021).

Diversification of the adhesin repertoire within a strain can arise from a variety of molecular mechanisms. For example, chimeric proteins generated through recombination between Als family members or between an Als protein's N terminal effector domain and an Hyr/Iff protein's repeat region have been shown (Butler et al. 2009; Zhao et al. 2011; Oh et al. 2019). Some of the adhesins with highly diverged central domains may have arisen in this manner (Fig. S10). Gene conversion between members of the same family can also drive the evolution of adhesin families within a species, as shown in *S. cerevisiae* and *C. glabrata* (Verstrepen et al. 2004; Marcet-Houben et al. 2022). Evidence of this in the Hil family was revealed in our analysis of recombination within the effector domain in *C. auris* (Fig. S8).

Special properties of the central domain in *C. auris* Hil1-Hil4 and related Hil proteins

A subset of Hil proteins represented by *C. auris* Hil1-Hil4 (Fig. 6A, B, vertical bar) stand out in that they are much longer on average and encode a large number of β-aggregation prone sequences compared with the rest of the family (Fig. 6B, D). Behind these properties is a conserved ~44 aa repeat unit containing a highly β-aggregation prone sequence ("GVVIVTT" and its variants) (Fig. 4B). β-aggregation prone sequences and the amyloid-like interaction they mediate have been extensively studied, especially in the Als protein family in *C. albicans*: they were experimentally shown to mediate aggregation (Otoo et al. 2008; Ramsook et al. 2010) and were crucial for forming protein clusters on cell surfaces known as nanodomains in response to physical tension or sheer forces (Alsteens et al. 2010; Lipke et al. 2012). Recently, they were also shown to mediate cell-cell *trans* interactions via homotypic protein binding (Dehullu et al. 2019; Ho et al. 2019). This may underlie biofilm formation and kin discrimination (Smukalla et al. 2008; Brückner et al. 2020; Lipke et al. 2021). Most known yeast adhesins, including the Als family proteins, encode between one and three β-aggregation prone sequences (Ramsook et al. 2010). *C. auris* Hil1-Hil4 and their close relatives are unusual in that they have as many as 50 such sequences, with each predicted by TANGO to have ~90% probability of aggregation, whereas the positive threshold for the algorithm is only >5% over 5-6 residues (Fernandez-Escamilla et al. 2004). The structural implications of the vast number of β-aggregation prone motifs may be that such tandem repeat domains are constitutively amyloid in nature, rather than requiring force or other stimuli as required by the Als proteins. The functional implications are unclear without the requisite experimental tests. However, we speculate that variations in protein length and β-aggregation potential resulting from the central domain divergence could directly impact the adhesion functions as previously suggested (Verstrepen et al. 2005; Boisramé et al. 2011; Lipke et al. 2012).

Structural predictions of the tandem repeat region in *C. auris* Hil1 and Hil2

Given the large number of ~44 aa repeats in the central domain of *C. auris* Hil1-Hil4 and the prediction that each repeat encodes 3-4 short consecutive β-strands (Fig 4B), we wondered what structural properties this region may have and how these features might contribute to the adhesion function. We explored this question using threading based structural prediction tools such as I-TASSER (Yang et al. 2015) and pDOMThreader (Lobley et al. 2009). For the tandem repeat region in the central domain of Hil1, I-TASSER identified (S)-layer protein (SLP) structures (e.g., RsaA from *C. crescentus*, SbsA and SbsC from *G. sterotherophilus*) as among the top structural analogs. These β-strand-rich structures are known to self-assemble to form a 2-dimensional array on the surface of bacteria, mediating a range of functions including adhesion to host cells in pathogens (Fagan and Fairweather 2014). pDOMThreader analyses of the central domains in Hil1 and Hil2 identified a different set of templates, namely bacterial self-associating proteins including Ag43a from uropathogenic *E. coli*, pertactin from *B. pertussis* and the *H. influenzae* hap adhesin. Interestingly, these proteins have β-helical structures like the Hyphal_reg_CWP domain, with the β-helices being involved in cell-cell interaction via an interface along the long solenoidal axis for homotypic interactions, and mediating bacterial clumping (Heras et al. 2014) and lead to biofilm formation in *H. influenzae* (Meng et al. 2011). We speculate that the long repeat regions in Hil1 and Hil2 may similarly mediate cell-cell interactions in *C. auris*.

The possibility that the central domains in Hil1 and Hil2 form a β-helix is interesting in that β-helix is one of the commonly described structural motifs in functional amyloids, e.g., HET-s from the fungus *Podospora anserina* (Wasmer et al. 2008). Such a solenoid-type amyloid is distinguished from other amyloid types in that the β-helices formed by repeats within the same protein, rather than among distinct monomeric proteins, are suggested to be stabilized not only by polar zippers and hydrophobic contacts, but also by electrostatic interactions between the alternating β-strands (Willbold et al. 2021). Other examples of amyloid forming proteins with a predicted β-helix structure include the imperfect repeat domain in the human premelanosome protein Pmel17 (Louros et al. 2016) and the extracellular curli proteins of Enterobacteriaceae that are involved in biofilm formation and adhesion to host cells (Shewmaker et al. 2009). The proposed solenoidal structure of the central domain of Hil1-Hil4 like proteins, if true, would have two significant implications. First, it confers the necessary rigidity and extended conformation required for cell wall anchored adhesins to extend into the surrounding extracellular milieu.

Second, the numerous β-strand rich repeats each containing a highly amyloid prone heptameric sequence, and capable of wrapping into a solenoidal shaped stack, is likely to substantially reduce the rate-limiting nucleation step, which limits the formation of, e.g., an Aβ amyloid fiber. This would allow the formation of extracellular extensions at low protein concentrations without the need for an extensive fiber lengthening process via the incorporation of additional monomeric units. Finally, the observation of solenoid-mediated intercellular interactions in the Hap adhesins suggests that Hil proteins may likewise have a biofilm related function.

Genomic context

As reported by (Muñoz et al. 2021), we found that the Hil family genes are preferentially located near chromosomal ends in *C. auris* and also in other species examined (Fig. 7). This is similar to previous findings for the Flo and Epa families (Teunissen and Steensma 1995; De Las Peñas et al. 2003; Xu et al. 2020; Xu et al. 2021), as well as the Als genes in some species (Oh et al. 2021). This location bias of the Hil and other adhesin families is likely a key mechanism for their dynamic expansion and sequence evolution via ectopic recombination (Anderson et al. 2015) and by Break-Induced Replication (Bosco and Haber 1998; Sakofsky and Malkova 2017; Xu et al. 2021). Another potential consequence of the Hil family genes being located in subtelomeres is that they may be subject to epigenetic silencing as an additional regulatory mechanism, which can be derepressed in response to stress (Ai et al. 2002). Such epigenetic regulation of the adhesin genes was found to generate cell surface heterogeneity in *S. cerevisiae* (Halme et al. 2004) and lead to hyperadherent phenotypes in *C. glabrata* (Castaño et al. 2005).

Concluding remarks

To address the lack of candidate adhesins in *C. auris*, we identified and characterized the Hyr/Iff-like (Hil) family in this species and all yeasts. Based on our results, we hypothesize that expansion and diversification of adhesin gene families is a key step towards the evolution of fungal pathogenesis and that variation in the adhesin repertoire contributes to within and between species differences in the adhesive and virulence properties. Future experimental tests of these hypotheses will be important biologically for improving our understanding of the fungal adhesin repertoire, biotechnologically for inspiring additional nanomaterials, and biomedically for advancing the development of *C. auris*-directed therapeutics.

Materials and Methods

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Bin Z. He (bin-he@uiowa.edu).

Data and code availability

All raw data and code for generating the intermediate and final results are available at the GitHub repository at https://github.com/binhe-lab/C037-Cand-auris-adhesin. Upon publication, this repository will be digitally archived with Zenodo and a DOI will be minted and provided to ensure reproducibility.

Software and algorithms list

NAME	REFERENCE	WEB OR DOWNLOAD URL
AlphaFold2	(Jumper et al. 2021)	https://github.com/sokrypton/ColabFold (links to DeepMind Google Colab Notebook)
BLAST+ v2.12.0	(Camacho et al. 2009)	https://blast.ncbi.nlm.nih.gov/
ClipKit	(Steenwyk et al. 2020)	https://github.com/JLSteenwyk/ClipKIT
Clustal Omega v1.2.4	(Sievers et al. 2011)	http://www.clustal.org/omega/
Custom R, Python and shell scripts	This study	https://github.com/binhe-lab/C037-Cand-auris-adhesin
DALI	(Holm 2022)	http://ekhidna2.biocenter.helsinki.fi/dali/
EMBOSS v6.6.0.0	(Rice et al. 2000)	http://emboss.open-bio.org/
FungalRV	(Chaudhuri et al. 2011)	http://fungalrv.igib.res.in/
GeneRax v2.0.1	(Morel et al. 2020)	https://github.com/BenoitMorel/GeneRax
HmmerWeb (hmmscan)	(Potter et al. 2018)	https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan
Jalview v2.11	(Waterhouse et al. 2009)	https://www.jalview.org/
JDotter	(Brodie et al. 2004)	https://4virology.net/virology-ca-tools/jdotter/
NetNGlyc v1.0	(Gupta and Brunak 2002)	https://services.healthtech.dtu.dk/service.php?NetNGlyc-1.0

NAME	REFERENCE	WEB OR DOWNLOAD URL
NetOGlyc v4.0	(Steentoft et al. 2013)	https://services.healthtech.dtu.dk/service.php?NetOGlyc-4.0
PAL2NAL.pl	(Suyama et al. 2006)	http://www.bork.embl.de/pal2nal/
PAML v4.9e	(Yang 2007)	http://abacus.gene.ucl.ac.uk/software/paml.html
PredGPI	(Pierleoni et al. 2008)	http://gpcr.biocomp.unibo.it/predgpi/
PSIPred	(Buchan and Jones 2019)	http://bioinf.cs.ucl.ac.uk/psipred/
PyMol v2.5.2	(Schrödinger, LLC 2021)	https://pymol.org/
R package - ggtree v3.2.1	(Yu 2020)	https://github.com/YuLab-SMU/ggtree
R package - phylolm	(Ho and Ané 2014)	https://cran.r-project.org/web/packages/phylolm/index.html
R package - rentrez v1.2.3	(Winter 2017)	https://github.com/ropensci/rentrez
R package - treeio v1.18.1	(Wang et al. 2020)	https://github.com/YuLab-SMU/treeio
R v4.1.0	(R Core Team)	https://cran.r-project.org/
RAxML v8.0.0	(Stamatakis 2014)	https://cme.h-its.org/exelixis/web/software/raxml/
RAxML-NG v1.1.0	(Kozlov et al. 2019)	https://github.com/amkozlov/raxml-ng
RStudio v1.4	(RStudio Team 2021)	https://www.rstudio.com/
SignalP 6.0	(Teufel et al. 2022)	http://www.cbs.dtu.dk/services/SignalP/
TANGO v2.3.1	(Fernandez-Escamilla et al. 2004)	http://tango.crg.es/
XSTREAM	(Newman and Cooper 2007)	https://amnewmanlab.stanford.edu/xstream/download.jsp

METHOD DETAILS

Identify Hyr/Iff-like (Hil) family homologs in yeasts and beyond

To identify the Hil family proteins in yeasts and beyond, we used the Hyphal_reg_CWP domain sequence from three distantly related Hil homologs as queries, namely, *C. albicans* Hyr1 (XP_722183.2), *C. auris* Hil1 (XP_028889033) and *C. glabrata* CAGL0E06600g (XP_722183.2). We performed BLASTP searches in the RefSeq protein database with an *E*-value cutoff of $1\text{x}10^{-5}$, a minimum query coverage of 50% and with the low complexity filter on. All hits were from Ascomycota (yeasts) and all but one were from the Saccharomycetes class (budding yeast). A single hit was found in the fission yeast *Schizosacchromyces cryophilus*. Using that hit as the query, we searched all fission yeasts in the nr protein database, with a relaxed E-value cutoff of 10^{-3} and identified no additional hits. We thus excluded that one hit from downstream analyses. To supplement the RefSeq database, which lacks some yeast species such as those in the Nakaseomyces genus, we searched the Genome Resources for Yeast Chromosomes (GRYC, http://gryc.inra.fr/). Using the same criteria, we recovered 16 additional sequences. To allow for gene tree and species tree reconciliation, we excluded three species that are not part of the 322 species yeast phylogeny (Shen et al. 2018) and not a member of the Multidrug-Resistant clade (Muñoz et al. 2018). Further details, including additional quality control steps taken to ensure that the homolog sequences are accurate and complete, can be found in Text S1. In total, we curated a list of 215 Hil family homologs from 32 species.

Gene family enrichment analysis

To determine if the Hil family is enriched in the pathogenic yeasts, we performed two analyses. In the first analysis, we divided the species into pathogens vs low-pathogenic potential groups and performed a t-test with unequal variance (also known as Welch's test) as well as a non-parametric Mann-Whitney U test to compare the Hil family size in the two groups. For both tests, we used either the total size of the family, or the number of putative adhesins as the random variable, and the results were consistent. We excluded homologs from *M. bicuspidata* because 10 of its 29 Hil family proteins were annotated as incomplete in the RefSeq protein database, and also because as a parasite of freshwater crustaceans, it does not fit into either the human pathogen or the low-pathogenic potential group. *S. cerevisiae* was included in the comparison as an example of species with zero members of the Hil family. We chose *S. cerevisiae* because we could be confident about its lack of a Hil family homolog thanks to its well assembled and well annotated genome.

In the second test, we used phylogenetic logistic regression (Ives and Garland 2010) to account for the phylogenetic relatedness between species. We used the `phyloglm` function in the `phylolm` package in R, with {method = "logistic_IG10", btol = 50, boot = 100}. The species tree, including the topology and branch lengths, were based on the 322 species phylogeny from (Shen et al. 2018), supplemented by the phylogenetic relationship for the MDR clade based on (Muñoz et al. 2018). The *P*-values based on phylogenetically specified residual correlations were reported.

Phylogenetic analysis of the Hil family and inference of gene duplications and losses

To infer the evolutionary history of the Hil family, we reconstructed a maximum-likelihood tree based on the alignment of the Hyphal_reg_CWP domain. First, we used hmmscan (HmmerWeb version 2.41.2) to identify the location of the Hyphal_reg_CWP domain in each Hil homolog. We used the “envelope boundaries” to define the domain in each sequence, and then aligned their amino acid sequences using Clustal Omega with the parameter {--iter=5}. We then trimmed the alignment using ClipKit with its default smart-gap trimming mode (Steenwyk et al. 2020). RAxML-NG v1.1.0 was run in the MPI mode with the following parameters on the alignment: “raxml-ng-mpi --all --msa INPUT --model LG+G --seed 123 --bs-trees autoMRE”. The resulting tree was corrected using GeneRax, which seeks to maximize the joint likelihood of observing the alignment given the gene family tree (GFT) and observing the GFT given the species phylogeny, using the parameter {--rec-model UndatedDL}. The species tree used is the same as the one used for the phylogenetic logistic regression above. In addition to correcting the gene family tree, GeneRax also reconciled it with the species tree and inferred duplication and loss event counts on each branch. Tree annotation and visualization were done in R using the treeio and ggtree packages (Wang et al. 2020; Yu 2020).

To infer the phylogenetic tree for the Hil family homologs in various *C. auris* strains and infer gains and losses within species, we identified orthologs of the *HIL* genes in representative strains from the four major clades of *C. auris* (B8441, B11220, B11221, B11243) (Muñoz et al. 2018). Orthologs from two MDR species, *C. haemuloni* and *C. pseudohaemulonis*, and from *D. hansenii* were included to help root the tree. The gene tree was constructed as described above. To root the tree, we first inferred a gene tree without the outgroup (*D. hansenii*) sequences in the alignment. Then, the full alignment with the outgroup sequences along with the gene tree from the first step were provided to RAxML to run the Evolutionary Placement Algorithm (EPA) algorithm (Berger et al. 2011), which identified a unique root location. To reconcile the gene tree with the species tree, we performed maximum likelihood-based gene tree correction using GeneRax (v2.0.1) with the parameters: {--rec-model UndatedDL}. The species tree was based on (Muñoz et al. 2018).

Prediction for fungal adhesins and adhesin-related sequence features

1) The potential of Hil homologs encoding fungal adhesins was assessed using FungalRV, a Support Vector Machine-based fungal adhesin predictor (Chaudhuri et al. 2011). Proteins passing the recommended cutoff of 0.511 were considered positive. **2)** Signal Peptide was predicted using the SignalP 6.0 server, with the "organism group" set to Eukarya. The server reported the proteins that had predicted signal peptides. No further filtering was done. **3)** GPI-anchor was predicted using PredGPI using the General Model. Proteins with a false positive rate of 0.01 or less were considered as containing a GPI-anchor. **4)** Tandem repeats were identified using XSTREAM with the following parameters: {-i.7 -I.7 -g3 -e2 -L15 -z -Asub.txt -B -O}, where the "sub.txt" was provided by the software package. **5)** β-aggregation prone sequences were predicted using TANGO v2.3.1 with the following parameters: {ct="N" nt="N" ph="7.5" te="298" io="0.1" tf="0" stab="-10" conc="1" seq="SEQ"}. **6)** Serine and Threonine content in proteins were quantified using `freak` from the EMBOSS suite, with a sliding window of 100 or 50 aa and a step size of 10 aa. To compare with proteome-wide distribution of Ser/Thr frequency, the protein sequences for *C. albicans* (SC5314), *C. glabrata* (CBS138) and *C. auris* (B11221) were downloaded from NCBI Assembly database (IDs in Table S7) and the frequency of serine and threonine residues were counted for each protein. **7)** O-linked and N-linked glycosylations were predicted using NetOGlyc (v4.0) and NetNGlyc (v1.0) servers.

Structural prediction and visualization for the Hyphal_reg_CWP domain

To perform structural predictions using AlphaFold2, we used the Google Colab notebook (https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb) authored by the DeepMind team. This is a reduced version of the full AlphaFold version 2 in that it searches a selected portion of the environmental BFD database, and doesn't use templates. The Amber relaxation step is included, and no other parameters other than the input sequences are required. DALI was used to search for similar structures in the PDB50 database. Model visualization and annotation were done in PyMol v2.5.2. Secondary structure prediction for *C. auris* Hil1's central domain was performed using PSIPred.

Dotplot

To determine the self-similarity and similarity between the eight *C. auris* Hil proteins, we made dotplots using JDotter (Brodie et al. 2004). The window size and contrast settings were labeled in the legends for the respective plots. The self-alignment for *C. auris* Hil1 tandem repeats was visualized using Jalview v2.11.

Identification of intraspecific tandem repeat copy number variations among *C. auris* strains

To identify polymorphisms in Hil1-Hil4 in diverse *C. auris* strains, we downloaded the genome sequences for the following strains from NCBI: Clade I - B11205, B13916; Clade II - B11220, B12043, B13463; Clade III - B11221, B12037, B12631, B17721; Clade IV - B11245, B12342 (Table S4). The amino acid sequences for Hil1-Hil4 from the strain B8441 were used as the query to search the nucleotide sequences of the above assemblies using TBLASTN, with the following parameters {-db_gencode 12 -evalue 1e-150 -max_hsps 2}. Orthologs in each strain were curated based on the BLAST hits to either the Hyphal_reg_CWP domain alone or the entire protein query. All Clade II strains had no hits for Hil1-Hil4. Several strains in Clade I, III and IV were found to lack one or more Hil proteins (Table S5). But upon further inspection, it was found that they had significant TBLASTN hits for part of the query, e.g., the central domain, and the hits were located at the end of a chromosome, suggesting the possibility of incomplete or misassembled sequences. Further experiments will be needed to determine if those *HIL* genes are present in those strains.

Estimation of dN/dS ratios and model comparisons

We used `codeml` in PAML (v4.9e) to perform evolutionary inferences on the Hyphal_reg_CWP domain in *C. auris*. We first used Clustal Omega to align the amino acid sequences for the Hyphal_reg_CWP domain from Hil1-Hil8 from *C. auris* similar to how we generated the multiple sequence alignment for all Hil proteins. A closely related outgroup (XP_018709340.1 from *M. bicuspidata*) was included to root the tree. We then generated a coding sequence alignment from the protein alignment using PAL2NAL (Suyama et al. 2006). We used GARD (Kosakovsky Pond et al. 2006) to analyze the coding sequence alignment to detect gene conversion events. The web service of GARD on datamonkey.org was run with the following parameters: {data type: nucleotide, run mode: normal, genetic code: yeast alternative nuclear, site-to-site rate variation: general discrete, rate classes: 3}. Based on the results, we identified two putatively non-recombining partitions, P1 = 1-414 and P2 = 697-981 (the numbers refer to the alignment columns). We then separately analyzed the two partitions in PAML. To test hypotheses about positive selection on a subset of the sites on all branches, we compared models M2a vs M1a, M8 vs M7 and M8a vs M8. The first 4 models were specified by: {seqtype = 1, CodonFreq = 1, model = 0, NSsites = 0,1,2,7,8, icode = 8, fix_kappa = 0, kappa = 2, fix_omega = 0, omega = 0.4, cleandata = 1}. The model M8a is additionally specified by { seqtype = 1, CodonFreq = 1, model = 0, NSsites = 8, fix_omega = 1 and omega = 1, cleandata = 1}. To test hypotheses for variable dN/dS on different branches (no variation across sites), we used {model = 0 or 1 or 2, NSsites = 0}, with the rest being the same as the site tests. Model = 0 specified the single ratio model, model = 1 the free ratio model and model = 2 the user-defined model. For the user-defined model, we first used estimates from the free ratio model to designate a set of branches with dN/dS > 10 as the foreground and then tested if their dN/dS was significantly different from the rest of the tree by comparing a two-ratio model with the single-ratio model. Since the results were significant, we further tested if the foreground dN/dS was significantly greater than 1, by comparing the two-ratio model to a constrained version of the model where omega was fixed at 1. For branch-site test, we used {model = 2, NSsites = 2, fix_omega = 0, omega = .4} as the alternative model and {model = 2, NSsites = 2, fix_omega = 1, omega = 1} as the null to test for positive selection on a subset of the sites on the foreground branches. Sites under positive selection were identified using the Bayes Empirical Bayes (BEB) procedure, with a posterior probability threshold of 0.99.

Chromosomal locations of Hil family genes

To compare the chromosomal locations of the Hil family genes to the background distribution, we selected eight species whose genomes were assembled to a chromosomal level and are not within a closely related group, including *C. albicans*, *D. hansenii*, *C. orthopsilosis*, *K. africana*, *K. lactis*, *N. dairenensis*, *C. auris* and *C. glabrata* (Table S7). We did not include some species, e.g., *C. dubliniensis*, to minimize statistical dependence due to shared ancestry. The RefSeq assembly for *C. auris* was included even though it was at a scaffold level because a recent study showed that seven of its longest scaffolds were chromosome-length, allowing the mapping of the scaffolds to chromosomes (Muñoz et al. 2021, Supplementary Table 1). To determine the chromosomal locations of the Hil homologs in these eight species, we used Rentrez v1.2.3 (Winter 2017) in R to retrieve their chromosome ID and coordinates. To calculate the background gene density on each chromosome, we downloaded the feature tables for the eight assemblies from the NCBI assembly database and calculated the location of each gene as its start coordinate divided by the chromosome length. To compare the chromosomal location of the Hil family genes to the genome background, we divided each chromosome into five equal-sized bins based on the physical distance to the nearest chromosomal end. We calculated the proportion of genes residing in each bin for the Hil family or for all protein coding genes. To determine if the two distributions differ significantly from one another, we performed a goodness-of-fit test using either a Log Likelihood Ratio (LLR) test or a Chi-Squared test, as implemented in the XNomial package in R (Engels 2015). The LLR test *P*-value was reported.

Acknowledgement

This study was inspired by work conducted during a Bioinformatics course (BIOL:4213) in Fall of 2019, whose semester theme was the recently published *C. auris* genomes. Graduate students, Lindsey Snyder and Rachel Smoak, were lead investigators on a group project centered on the genes containing the Hyphal_reg_CWP (PF11765) domain during that course, and were happy to be invited to continue working on these genes long after the semester ended. We would also like to thank members of the Gene Regulatory Evolution lab for discussion. We thank Drs. Yong Zhang (CAS Zoology Institute), Peter Lipke (Brooklyn College), John McCutcheon (ASU), Matthew Hahn (Indiana University), Rich Lenski (MSU), Josep Comeron (University of Iowa), Daniel Weeks (University of Iowa), Lois Hoyer (UIUC) and two other anonymous reviewers for providing useful suggestions and critical comments. Dr. Bin Z. He is supported by NIH R35GM137831. Dr. Jan Fassler is supported by NSF-IOS-1917169. Lindsey F. Snyder was supported by the NIH Predoctoral Training grant T32GM008629. Rachel A. Smoak is supported by an NSF Graduate Research Fellowship Program under Grant No. 1546595, with additional support through the NSF Division of Graduate Education under Grant No. 1633098.

Reference

Ai W, Bertram PG, Tsang CK, Chan TF, Zheng XFS. 2002. Regulation of subtelomeric silencing during stress response. *Mol. Cell* 10:1295–1305.

Alsteens D, Garcia MC, Lipke PN, Dufrêne YF. 2010. Force-induced formation and propagation of adhesion nanodomains in living fungal cells. *Proc. Natl. Acad. Sci. U. S. A.* 107:20744–20749.

Anderson MZ, Wigen LJ, Burrack LS, Berman J. 2015. Real-Time Evolution of a Subtelomeric Gene Family in Candida albicans. *Genetics* 200:907–919.

Bailey DA, Feldmann PJ, Bovey M, Gow NA, Brown AJ. 1996. The Candida albicans HYR1 gene, which is activated in response to hyphal development, belongs to a gene family encoding yeast cell wall proteins. *J. Bacteriol.* 178:5353–5360.

Bates S, de la Rosa JM, MacCallum DM, Brown AJP, Gow NAR, Odds FC. 2007. Candida albicans Iff11, a secreted protein required for cell wall structure and virulence. *Infect. Immun.* 75:2922–2928.

Berger SA, Krompass D, Stamatakis A. 2011. Performance, accuracy, and Web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst. Biol.* 60:291–302.

Boisramé A, Cornu A, Da Costa G, Richard ML. 2011. Unexpected role for a serine/threonine-rich domain in the Candida albicans Iff protein family. *Eukaryot. Cell* 10:1317–1330.

Bosco G, Haber JE. 1998. Chromosome break-induced DNA replication leads to nonreciprocal translocations and telomere capture. *Genetics* 150:1037–1047.

Bradley PH, Nayfach S, Pollard KS. 2018. Phylogeny-corrected identification of microbial gene families relevant to human gut colonization. *PLOS Comput. Biol.* 14:e1006242.

Brodie R, Roper RL, Upton C. 2004. JDotter: a Java interface to multiple dotplots generated by dotter. *Bioinforma. Oxf. Engl.* 20:279–281.

Brückner S, Schubert R, Kraushaar T, Hartmann R, Hoffmann D, Jelli E, Drescher K, Müller DJ, Oliver Essen L, Mösch H-U. 2020. Kin discrimination in social yeast is mediated by cell surface receptors of the Flo11 adhesin family.Brakhage AA, Barkai N, editors. *eLife* 9:e55587.

Buchan DWA, Jones DT. 2019. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.* 47:W402–W407.

Butler G, Rasmussen MD, Lin MF, Santos MAS, Sakthikumar S, Munro CA, Rheinbay E, Grabherr M, Forche A, Reedy JL, et al. 2009. Evolution of pathogenicity and sexual reproduction in eight Candida genomes. *Nature* 459:657–662.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

Casola C, Hahn MW. 2009. Gene conversion among paralogs results in moderate false detection of positive selection using likelihood methods. *J. Mol. Evol.* 68:679–687.

Castaño I, Pan S-J, Zupancic M, Hennequin C, Dujon B, Cormack BP. 2005. Telomere length control and transcriptional regulation of subtelomeric adhesins in Candida glabrata. *Mol. Microbiol.* 55:1246–1258.

CDC. 2019. Antibiotic resistance threats in the United States, 2019. *US Dep. Health Hum. Serv. CDC* [Internet]. Available from: https://stacks.cdc.gov/view/cdc/82532

Chaudhuri R, Ansari FA, Raghunandanan MV, Ramachandran S. 2011. FungalRV: adhesin prediction and immunoinformatics portal for human fungal pathogens. *BMC Genomics* 12:192.

De Las Peñas A, Pan S-J, Castaño I, Alder J, Cregg R, Cormack BP. 2003. Virulence-related surface glycoproteins in the yeast pathogen Candida glabrata are encoded in subtelomeric clusters and subject to RAP1- and SIR-dependent transcriptional silencing. *Genes Dev.* 17:2245–2258.

Dehullu J, Valotteau C, Herman-Bausier P, Garcia-Sherman M, Mittelviefhaus M, Vorholt JA, Lipke PN, Dufrêne YF. 2019. Fluidic Force Microscopy Demonstrates That Homophilic Adhesion by Candida albicans Als Proteins Is Mediated by Amyloid Bonds between Cells. *Nano Lett.* 19:3846–3853.

Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, de Montigny J, Marck C, Neuvéglise C, Talla E, et al. 2004. Genome evolution in yeasts. *Nature* 430:35–44.

Eberlein C, Nielly-Thibault L, Maaroufi H, Dubé AK, Leducq J-B, Charron G, Landry CR. 2017. The Rapid Evolution of an Ohnolog Contributes to the Ecological Specialization of Incipient Yeast Species. *Mol. Biol. Evol.* 34:2173–2186.

Engels B. 2015. XNomial: Exact Goodness-of-Fit Test for Multinomial Data with Fixed Probabilities. Available from: https://CRAN.R-project.org/package=XNomial

Fagan RP, Fairweather NF. 2014. Biogenesis and functions of bacterial S-layers. *Nat. Rev. Microbiol.* 12:211–222.

Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L. 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* 22:1302–1306.

Frank AT, Ramsook CB, Otoo HN, Tan C, Soybelman G, Rauceo JM, Gaur NK, Klotz SA, Lipke PN. 2010. Structure and Function of Glycosylated Tandem Repeats from Candida albicans Als Adhesins. *Eukaryot. Cell* 9:405–414.

Frieman MB, McCaffery JM, Cormack BP. 2002. Modular domain structure in the Candida glabrata adhesin Epa1p, a beta1,6 glucan-cross-linked cell wall protein. *Mol. Microbiol.* 46:479–492.

Fu Y, Luo G, Spellberg BJ, Edwards JE, Ibrahim AS. 2008. Gene overexpression/suppression analysis of candidate virulence factors of Candida albicans. *Eukaryot. Cell* 7:483–492.

Gabaldón T, Martin T, Marcet-Houben M, Durrens P, Bolotin-Fukuhara M, Lespinet O, Arnaise S, Boisnard S, Aguileta G, Atanasova R, et al. 2013. Comparative genomics of emerging pathogens in the Candida glabrata clade. *BMC Genomics* 14:623.

Gabaldón T, Naranjo-Ortíz MA, Marcet-Houben M. 2016. Evolutionary genomics of yeast pathogens in the Saccharomycotina. *FEMS Yeast Res.* 16.

Gordon JL, Armisén D, Proux-Wéra E, ÓhÉigeartaigh SS, Byrne KP, Wolfe KH. 2011. Evolutionary erosion of yeast sex chromosomes by mating-type switching accidents. *Proc. Natl. Acad. Sci.* 108:20024–20029.

de Groot PWJ, Bader O, de Boer AD, Weig M, Chauhan N. 2013. Adhesins in human fungal pathogens: glue with plenty of stick. *Eukaryot. Cell* 12:470–481.

Gupta R, Brunak S. 2002. Prediction of glycosylation across the human proteome and the correlation to protein function. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.*:310–322.

Hall SR, Becker CR, Duffy MA, Cáceres CE. 2010. Variation in resource acquisition and use among host clones creates key epidemiological trade-offs. *Am. Nat.* 176:557–565.

Halme A, Bumgarner S, Styles C, Fink GR. 2004. Genetic and Epigenetic Regulation of the FLO Gene Family Generates Cell-Surface Variation in Yeast. *Cell* 116:405–415.

Heras B, Totsika M, Peters KM, Paxman JJ, Gee CL, Jarrott RJ, Perugini MA, Whitten AE, Schembri MA. 2014. The antigen 43 structure reveals a molecular Velcro-like mechanism of autotransporter-mediated bacterial clumping. *Proc. Natl. Acad. Sci. U. S. A.* 111:457–462.

Ho L si T, Ané C. 2014. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. *Syst. Biol.* 63:397–408.

Ho V, Herman-Bausier P, Shaw C, Conrad KA, Garcia-Sherman MC, Draghi J, Dufrene YF, Lipke PN, Rauceo JM. 2019. An Amyloid Core Sequence in the Major Candida albicans Adhesin Als1p Mediates Cell-Cell Adhesion. *mBio* 10:e01766-19.

Holm L. 2022. Dali server: structural unification of protein families. *Nucleic Acids Res.*:gkac387.

Hoyer LL. 2001. The ALS gene family of Candida albicans. *Trends Microbiol.* 9:176–180.

HOYER LL, GREEN CB, OH S-H, ZHAO X. 2008. Discovering the Secrets of the Candida albicans Agglutinin-Like Sequence (ALS) Gene Family—a Sticky Pursuit. *Med. Mycol. Off. Publ. Int. Soc. Hum. Anim. Mycol.* 46:1–15.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 11:97–108.

Ives AR, Garland T. 2010. Phylogenetic logistic regression for binary dependent variables. *Syst. Biol.* 59:9–26.

Jenull S, Tscherner M, Kashko N, Shivarathri R, Stoiber A, Chauhan M, Petryshyn A, Chauhan N, Kuchler K. 2021. Transcriptome Signatures Predict Phenotypic Variations of Candida auris. *Front. Cell. Infect. Microbiol.* [Internet] 11. Available from: https://www.ncbi.nlm.nih.gov/labs/pmc/articles/PMC8079977/

Jiang H, Bao J, Xing Y, Li X, Chen Q. 2022. Comparative Genomic Analyses Provide Insight Into the Pathogenicity of Metschnikowia bicuspidata LNES0119. *Front. Microbiol.* 13:939141.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*:1–11.

Kajava AV, Steven AC. 2006. The turn of the screw: variations of the abundant beta-solenoid motif in passenger domains of Type V secretory proteins. *J. Struct. Biol.* 155:306–315.

Kean R, Delaney C, Sherry L, Borman A, Johnson EM, Richardson MD, Rautemaa-Richardson R, Williams C, Ramage G. 2018. Transcriptome Assembly and Profiling of Candida auris Reveals Novel Insights into Biofilm-Mediated Resistance. *mSphere* [Internet] 3. Available from: https://msphere.asm.org/content/3/4/e00334-18

Kempf M, Cottin J, Licznar P, Lefrançois C, Robert R, Apaire-Marchais V. 2009. Disruption of the GPI protein-encoding gene IFF4 of Candida albicans results in decreased adherence and virulence. *Mycopathologia* 168:73–77.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006. Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm. *Mol. Biol. Evol.* 23:1891–1901.

Koteiche HA, Mchaourab HS. 1999. Folding pattern of the alpha-crystallin domain in alphaA-crystallin determined by site-directed spin labeling. *J. Mol. Biol.* 294:561–577.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinforma. Oxf. Engl.* 35:4453–4455.

Kuang MC, Hutchins PD, Russell JD, Coon JJ, Hittinger CT. 2016. Ongoing resolution of duplicate gene functions shapes the diversification of a metabolic network. *eLife* 5:e19027.

Kwon YJ, Shin JH, Byun SA, Choi MJ, Won EJ, Lee D, Lee SY, Chun S, Lee JH, Choi HJ, et al. 2019. Candida auris Clinical Isolates from South Korea: Identification, Antifungal Susceptibility, and Genotyping. *J. Clin. Microbiol.* 57:e01624-18.

Lamoth F, Lockhart SR, Berkow EL, Calandra T. 2018. Changes in the epidemiological landscape of invasive candidiasis. *J. Antimicrob. Chemother.* 73:i4–i13.

Levy A, Salas Gonzalez I, Mittelviefhaus M, Clingenpeel S, Herrera Paredes S, Miao J, Wang K, Devescovi G, Stillman K, Monteiro F, et al. 2017. Genomic features of bacterial adaptation to plants. *Nat. Genet.* 50:138–150.

Linder T, Gustafsson CM. 2008. Molecular phylogenetics of ascomycotal adhesins—A novel family of putative cell-surface adhesive proteins in fission yeasts. *Fungal Genet. Biol.* 45:485–497.

Lipke PN. 2018. What We Do Not Know about Fungal Cell Adhesion Molecules. *J. Fungi Basel Switz.* 4.

Lipke PN, Garcia MC, Alsteens D, Ramsook CB, Klotz SA, Dufrêne YF. 2012. Strengthening relationships: amyloids create adhesion nanodomains in yeasts. *Trends Microbiol.* 20:59–65.

Lipke PN, Mathelié-Guinlet M, Viljoen A, Dufrêne YF. 2021. A New Function for Amyloid-Like Interactions: Cross-Beta Aggregates of Adhesins form Cell-to-Cell Bonds. *Pathogens* 10:1013.

Lobley A, Sadowski MI, Jones DT. 2009. pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinforma. Oxf. Engl.* 25:1761–1767.

Lockhart SR, Etienne KA, Vallabhaneni S, Farooqi J, Chowdhary A, Govender NP, Colombo AL, Calvo B, Cuomo CA, Desjardins CA, et al. 2017. Simultaneous Emergence of Multidrug-Resistant Candida auris on 3 Continents Confirmed by Whole-Genome Sequencing and Epidemiological Analyses. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* 64:134–140.

Louros NN, Baltoumas FA, Hamodrakas SJ, Iconomidou VA. 2016. A β-solenoid model of the Pmel17 repeat domain: insights to the formation of functional amyloid fibrils. *J. Comput. Aided Mol. Des.* 30:153–164.

Luo G, Ibrahim AS, Spellberg B, Nobile CJ, Mitchell AP, Fu Y. 2010. Candida albicans Hyr1p Confers Resistance to Neutrophil Killing and Is a Potential Vaccine Target. *J. Infect. Dis.* 201:1718–1728.

Marcet-Houben M, Alvarado M, Ksiezopolska E, Saus E, de Groot PWJ, Gabaldón T. 2022. Chromosome-level assemblies from diverse clades reveal limited structural and gene content variation in the genome of Candida glabrata. *BMC Biol.* 20:226.

Mefford HC, Trask BJ. 2002. The complex structure and dynamic evolution of human subtelomeres. *Nat. Rev. Genet.* 3:91–102.

Meng G, Spahich N, Kenjale R, Waksman G, St Geme JW. 2011. Crystal structure of the Haemophilus influenzae Hap adhesin reveals an intercellular oligomerization mechanism for bacterial aggregation. *EMBO J.* 30:3864–3874.

Morel B, Kozlov AM, Stamatakis A, Szöllősi GJ. 2020. GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss. *Mol. Biol. Evol.* 37:2763–2774.

Muñoz JF, Gade L, Chow NA, Loparev VN, Juieng P, Berkow EL, Farrer RA, Litvintseva AP, Cuomo CA. 2018. Genomic insights into multidrug-resistance, mating and virulence in Candida auris and related emerging species. *Nat. Commun.* 9:5346.

Muñoz JF, Welsh RM, Shea T, Batra D, Gade L, Howard D, Rowe LA, Meis JF, Litvintseva AP, Cuomo CA. 2021. Clade-specific chromosomal rearrangements and loss of subtelomeric adhesins in Candida auris. *Genetics* [Internet]. Available from: https://doi.org/10.1093/genetics/iyab029

Newman AM, Cooper JB. 2007. XSTREAM: A practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics* 8:382.

Nozawa M, Suzuki Y, Nei M. 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc. Natl. Acad. Sci.* 106:6700–6705.

Oh S-H, Isenhower A, Rodriguez-Bobadilla R, Smith B, Jones J, Hubka V, Fields C, Hernandez A, Hoyer LL. 2020. Pursuing Advances in DNA Sequencing Technology to Solve a Complex Genomic Jigsaw Puzzle: The Agglutinin-Like Sequence (ALS) Genes of Candida tropicalis. *Front. Microbiol.* 11:594531.

Oh S-H, Schliep K, Isenhower A, Rodriguez-Bobadilla R, Vuong VM, Fields CJ, Hernandez AG, Hoyer LL. 2021. Using Genomics to Shape the Definition of the Agglutinin-Like Sequence (ALS) Family in the Saccharomycetales. *Front. Cell. Infect. Microbiol.* 11:794529.

Oh S-H, Smith B, Miller AN, Staker B, Fields C, Hernandez A, Hoyer LL. 2019. Agglutinin-Like Sequence (ALS) Genes in the Candida parapsilosis Species Complex: Blurring the

Boundaries Between Gene Families That Encode Cell-Wall Proteins. *Front. Microbiol.* 10:781.

Otoo HN, Lee KG, Qiu W, Lipke PN. 2008. Candida albicans Als Adhesins Have Conserved Amyloid-Forming Sequences. *Eukaryot. Cell* 7:776–782.

Pierleoni A, Martelli PL, Casadio R. 2008. PredGPI: a GPI-anchor predictor. *BMC Bioinformatics* 9:392.

Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. 2018. HMMER web server: 2018 update. *Nucleic Acids Res.* 46:W200–W204.

Qian W, Zhang JG. 2014. Genomic evidence for adaptation by gene duplication. *Genome Res.*:gr.172098.114.

R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing Available from: https://www.R-project.org

Ramsook CB, Tan C, Garcia MC, Fung R, Soybelman G, Henry R, Litewka A, O'Meally S, Otoo HN, Khalaf RA, et al. 2010. Yeast cell adhesion molecules have functional amyloid-forming sequences. *Eukaryot. Cell* 9:393–404.

Rauceo JM, De Armond R, Otoo H, Kahn PC, Klotz SA, Gaur NK, Lipke PN. 2006. Threonine-rich repeats increase fibronectin binding in the Candida albicans adhesin Als5p. *Eukaryot. Cell* 5:1664–1673.

Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS. 2000. Parallel evolution of virulence in pathogenic Escherichia coli. *Nature* 406:64–67.

Reithofer V, Fernández-Pereira J, Alvarado M, de Groot P, Essen L-O. 2021. A novel class of Candida glabrata cell wall proteins with β-helix fold mediates adhesion in clinical isolates. *PLoS Pathog.* 17:e1009980.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet. TIG* 16:276–277.

Richard ML, Plaine A. 2007. Comprehensive Analysis of Glycosylphosphatidylinositol-Anchored Proteins in Candida albicans. *Eukaryot. Cell* 6:119–133.

Rosiana S, Zhang L, Kim GH, Revtovich AV, Uthayakumar D, Sukumaran A, Geddes-McAlister J, Kirienko NV, Shapiro RS. 2021. Comprehensive genetic analysis of adhesin proteins and their role in virulence of Candida albicans. *Genetics* [Internet]. Available from: https://doi.org/10.1093/genetics/iyab003

RStudio Team. 2021. RStudio: Integrated Development Environment for R. Boston, MA: RStudio, PBC Available from: http://www.rstudio.com/

Sakofsky CJ, Malkova A. 2017. Break induced replication in eukaryotes: mechanisms, functions, and consequences. *Crit. Rev. Biochem. Mol. Biol.* 52:395–413.

Schrödinger, LLC. 2021. The PyMOL Molecular Graphics System, Version 2.5.2.

Sequeira S, Kavanaugh D, MacKenzie DA, Šuligoj T, Walpole S, Leclaire C, Gunning AP, Latousakis D, Willats WGT, Angulo J, et al. 2018. Structural basis for the role of serine-rich repeat proteins from Lactobacillus reuteri in gut microbe–host interactions. *Proc. Natl. Acad. Sci.* 115:E2706–E2715.

Shen X-X, Opulente DA, Kominek J, Zhou X, Steenwyk JL, Buh KV, Haase MAB, Wisecaver JH, Wang M, Doering DT, et al. 2018. Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* [Internet]. Available from: http://www.sciencedirect.com/science/article/pii/S0092867418313321

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7:539.

Singh S, Uppuluri P, Mamouei Z, Alqarihi A, Elhassan H, French S, Lockhart SR, Chiller T, Jr JEE, Ibrahim AS. 2019. The NDV-3A vaccine protects mice from multidrug resistant Candida auris infection. *PLOS Pathog.* 15:e1007460.

Smukalla S, Caldara M, Pochet N, Beauvais A, Guadagnini S, Yan C, Vinces MD, Jansen A, Prevost MC, Latgé J-P, et al. 2008. FLO1 is a variable green beard gene that drives biofilm-like cooperation in budding yeast. *Cell* 135:726–737.

Srivastava V, Singla RK, Dubey AK. 2018. Emerging virulence, drug resistance and future anti-fungal drugs for Candida pathogens. *Curr. Top. Med. Chem.*

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

Stamler R, Kappé G, Boelens W, Slingsby C. 2005. Wrapping the α-Crystallin Domain Fold in a Chaperone Assembly. *J. Mol. Biol.* 353:68–79.

Steentoft C, Vakhrushev SY, Joshi HJ, Kong Y, Vester-Christensen MB, Schjoldager KT-BG, Lavrsen K, Dabelsteen S, Pedersen NB, Marcos-Silva L, et al. 2013. Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.* 32:1478–1488.

Steenwyk JL, Buida TJ, Li Y, Shen X-X, Rokas A. 2020. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol.* 18:e3001007.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34:W609-612.

Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G, Nielsen H. 2022. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* 40:1023–1025.

Teunissen AW, Steensma HY. 1995. Review: the dominant flocculation genes of Saccharomyces cerevisiae constitute a new subtelomeric gene family. *Yeast Chichester Engl.* 11:1001–1013.

Uppuluri P, Lin L, Alqarihi A, Luo G, Youssef EG, Alkhazraji S, Yount NY, Ibrahim BA, Bolaris MA, Edwards JE, et al. 2018. The Hyr1 protein from the fungus Candida albicans is a cross kingdom immunotherapeutic target for Acinetobacter bacterial infection. *PLoS Pathog.* [Internet] 14. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5963808/

Verstrepen KJ, Jansen A, Lewitter F, Fink GR. 2005. Intragenic tandem repeats generate functional variability. *Nat. Genet.* 37:986–990.

Verstrepen KJ, Reynolds TB, Fink GR. 2004. Origins of variation in the fungal cell surface. *Nat. Rev. Microbiol.* 2:533–540.

Wang L-G, Lam TT-Y, Xu S, Dai Z, Zhou L, Feng T, Guo P, Dunn CW, Jones BR, Bradley T, et al. 2020. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Mol. Biol. Evol.* 37:599–603.

Wasmer C, Lange A, Melckebeke HV, Siemer AB, Riek R, Meier BH. 2008. Amyloid Fibrils of the HET-s(218–289) Prion Form a β Solenoid with a Triangular Hydrophobic Core. *Science* [Internet]. Available from: https://www.science.org/doi/abs/10.1126/science.1151839

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinforma. Oxf. Engl.* 25:1189–1191.

Welsh RM, Sexton DJ, Forsberg K, Vallabhaneni S, Litvintseva A. 2019. Insights into the Unique Nature of the East Asian Clade of the Emerging Pathogenic Yeast Candida auris. *J. Clin. Microbiol.* 57:e00007-19.

Wilkins M, Zhang N, Schmid J. 2018. Biological Roles of Protein-Coding Tandem Repeats in the Yeast Candida Albicans. *J. Fungi* 4:78.

Willaert R. 2018. Adhesins of Yeasts: Protein Structure and Interactions. *J. Fungi* 4:119.

Willbold D, Strodel B, Schröder GF, Hoyer W, Heise H. 2021. Amyloid-type Protein Aggregation and Prion-like Properties of Amyloids. *Chem. Rev.* 121:8285–8307.

Winter DJ. 2017. rentrez: an R package for the NCBI eUtils API. *R J.* 9:520–526.

Xie X, Qiu W-G, Lipke PN. 2011. Accelerated and adaptive evolution of yeast sexual adhesins. *Mol. Biol. Evol.* 28:3127–3137.

Xu Z, Green B, Benoit N, Schatz M, Wheelan S, Cormack B. 2020. De novo genome assembly of Candida glabrata reveals cell wall protein complement and structure of dispersed tandem repeat arrays. *Mol. Microbiol.*

Xu Z, Green B, Benoit N, Sobel JD, Schatz MC, Wheelan S, Cormack BP. 2021. Cell wall protein variation, break-induced replication, and subtelomere dynamics in Candida glabrata. *Mol. Microbiol.*

Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. 2015. The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* 12:7–8.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15:568–573.

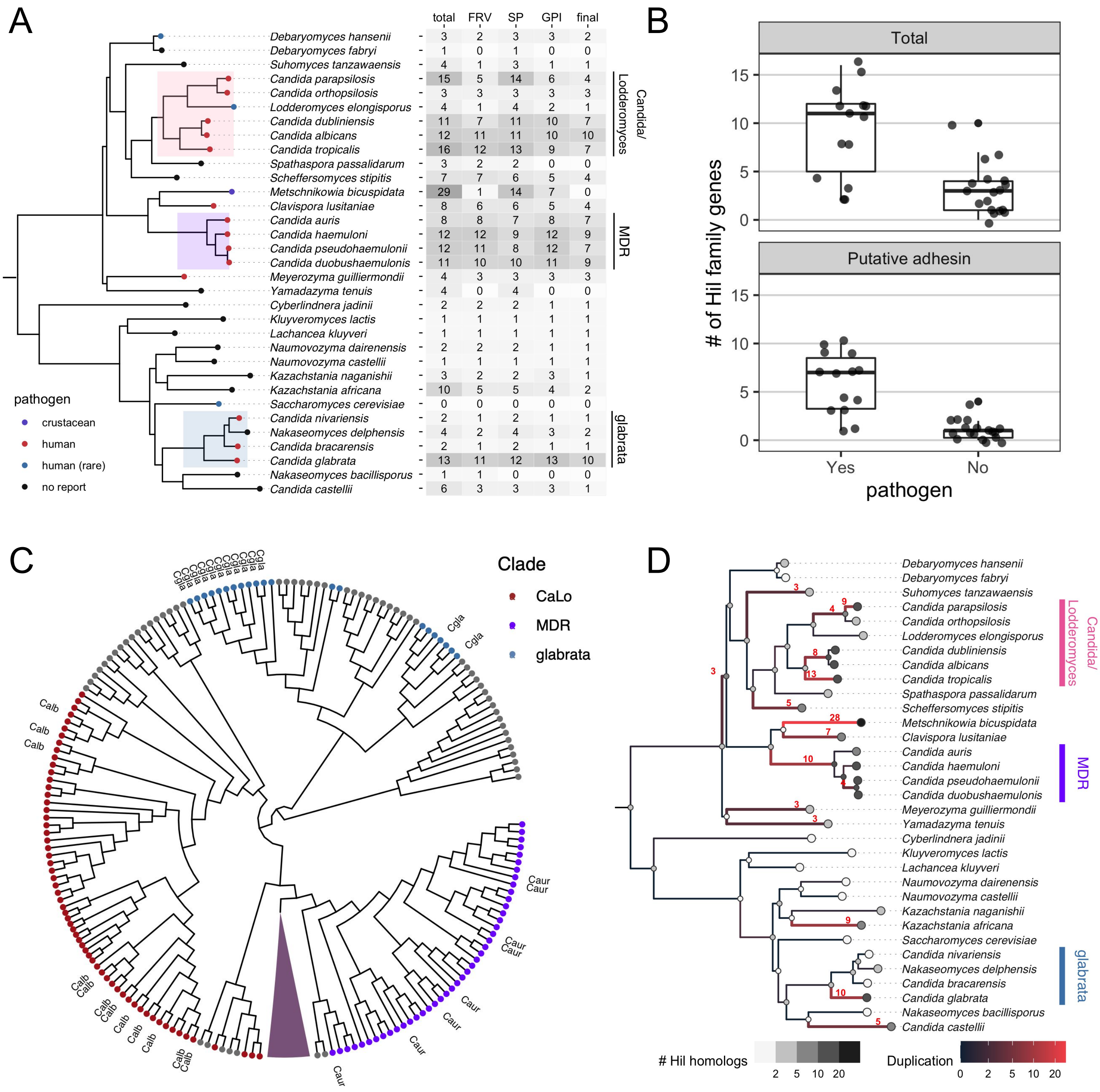
Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* 24:1586–1591.

Yu G. 2020. Using ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinforma.* 69:e96.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18:292–298.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Mol. Biol. Evol.* 22:2472–2479.

Zhao X, Oh S-H, Coleman DA, Hoyer LL. 2011. ALS51, a newly discovered gene in the Candida albicans ALS family, created by intergenic recombination: analysis of the gene and protein, and implications for evolution of microbial gene families. *FEMS Immunol. Med. Microbiol.* 61:245–257.



	total	FRV	SP	GPI	final
Debaryomyces hansenii	3	2	3	3	2
Debaryomyces fabryi	1	0	1	0	0
Suhomyces tanzawaensis	4	1	3	1	1
Candida parapsilosis	15	5	14	6	4
Candida orthopsilosis	3	3	3	3	3
Lodderomyces elongisporus	4	1	4	2	1
Candida dubliniensis	11	7	11	10	7
Candida albicans	12	11	11	10	10
Candida tropicalis	16	12	13	9	7
Spathaspora passalidarum	3	2	2	0	0
Scheffersomyces stipitis	7	7	6	5	4
Metschnikowia bicuspidata	29	1	14	7	0
Clavispora lusitaniae	8	6	6	5	4
Candida auris	8	8	7	8	7
Candida haemuloni	12	12	9	12	9
Candida pseudohaemulonii	12	11	8	12	7
Candida duobushaemulonis	11	10	10	11	9
Meyerozyma guilliermondii	4	3	3	3	3
Yamadazyma tenuis	4	0	4	0	0
Cyberlindnera jadinii	2	2	2	1	1
Kluyveromyces lactis	1	1	1	1	1
Lachancea kluyveri	1	1	1	1	1
Naumovozyma dairenensis	2	2	2	1	1
Naumovozyma castellii	1	1	1	1	1
Kazachstania naganishii	3	2	2	3	1
Kazachstania africana	10	5	5	4	2
Saccharomyces cerevisiae	0	0	0	0	0
Candida nivariensis	2	1	2	1	1
Nakaseomyces delphensis	4	2	4	3	2
Candida bracarensis	2	1	2	1	1
Candida glabrata	13	11	12	13	10
Nakaseomyces bacillisporus	1	1	0	0	0
Candida castellii	6	3	3	3	1

Figure 1. Phylogenetic distribution of the yeast Hil family and its parallel expansion in independently derived pathogenic *Candida* species. Legend on the next page.

Figure 1. Phylogenetic distribution of the yeast Hil family and its parallel expansion in independently derived pathogenic *Candida* species. A) Species tree is based on the phylogeny for 332 yeast species from (Shen *et al.* 2018), except for three species in the MDR clade other than *C. auris*, whose phylogenetic relationships are based on (Muñoz *et al.* 2018). The tip colors show the pathogenic status of the species. The highlighted clades are enriched in known human pathogens. In the table, the first column shows the total number of Hil family homologs per species. The number of homologs that pass each of the three tests for determining their adhesin status are shown in the next three columns. FRV = FungalRV, SP = Signal Peptide and GPI = GPI-anchor. See Materials and Methods for details. The number of homologs passing all three tests is shown in the "final" columns. (B) Boxplots comparing the number of Hil homologs (upper) or the number of putative adhesins passing all three tests (lower) per species between known pathogens and low pathogenic-potential species. Individual species numbers are shown as dots on top of the boxplot. Homologs from *M. bicuspidata* were excluded (see text). Both comparisons are significant at a 0.005 level by either a t-test with unequal variance or Mann-Whitney U test. (C) Maximum likelihood tree based on the Hyphal_reg_CWP domain of the Hil family was constructed using RAxML-NG and corrected with GeneRax based on the species tree. The tree is shown as a cladogram. All 29 homologs in *M. bicuspidata* formed a single group, which is shown as a triangle in dark plum. Homologs from the species in the three highlighted clades in (A) are colored accordingly. CaLo = Candida/Lodderomyces. Homologs from *C. albicans*, *C. auris* and *C. glabrata* are labeled as Calb, Caur and Cgla, respectively. (D) Species tree showing the number of inferred duplication events on each branch. The gray colors of the tip and internal nodes represent the identified and inferred number of Hil homologs, respectively. The branch color shows the inferred number of duplication events, with 3 or more duplications also shown as a number next to the branch.

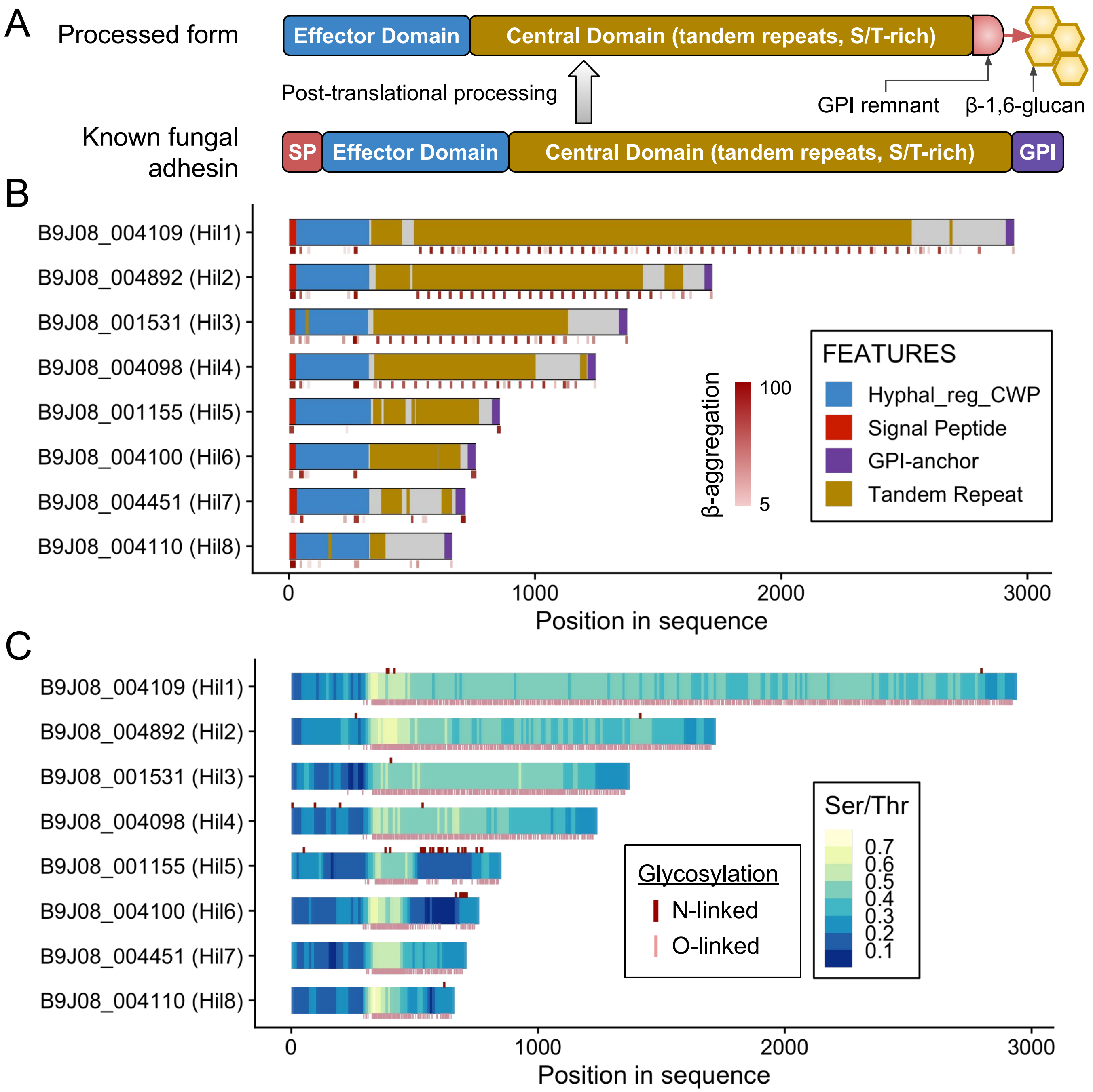


Figure 2. Domain architecture and adhesin-associated features of the *C. auris* Hil family. (A) Diagram depicting the domain organization of a typical yeast adhesin before and after the post-translational processing, adapted from (de Groot *et al.* 2013). (B) Domain features of the eight Hil proteins in *C. auris* (strain B8441). Gene IDs and names are labeled on the left. The short stripes below each diagram are the TANGO predicted β-aggregation prone sequences, with the intensity of the color corresponding to the score of the prediction. (C) Serine and Threonine (Ser/Thr) frequencies in each protein are plotted in 50 aa sliding windows with step size of 10 aa. N-linked and O-linked glycosylation sites were predicted by NetNGlyc 1.0 and NetOGlyc 4.0, respectively, and are shown as short ticks above and below each protein schematic.

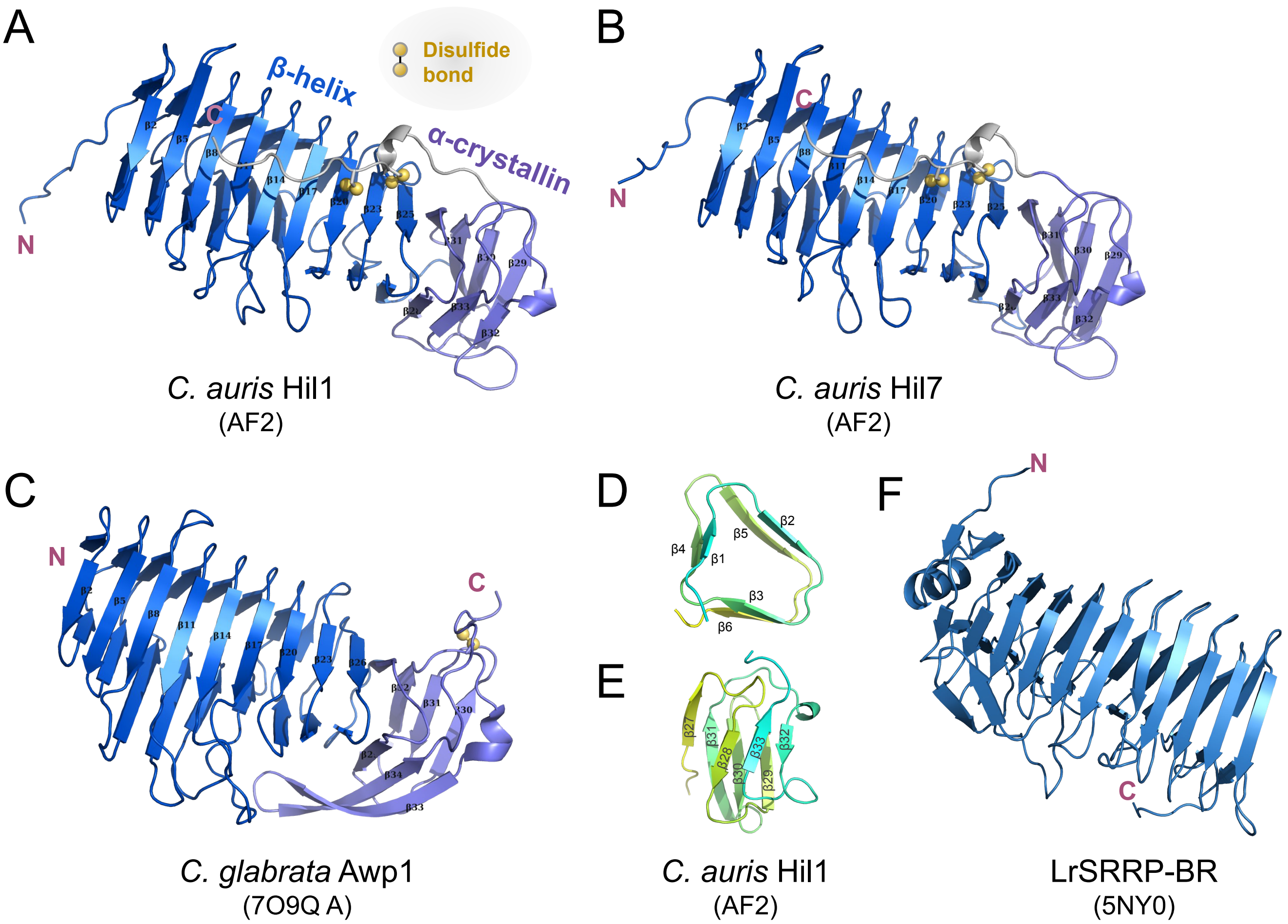


Figure 3. Predicted structures of the Hyphal_reg_CWP domain in two *C. auris* Hil proteins are similar to yeast and bacterial adhesins. (A) and (B) AlphaFold2 (AF2) predicted structures of the Hyphal_reg_CWP domains from *C. auris* Hil1 and Hil7, which consist of a β-helix followed by a α-crystallin domain, with the C-terminal loop linked to the β-helix via two disulfide bonds. (C) Crystal structure of the *C. glabrata* Awp1 effector domain, which is highly similar to *C. auris* Hil1 and Hil7, but with the disulfide bond in a different location. (D) cross-section of the first two rungs of the β-helix in (A), showing the three β-strands per rung. The cyan-to-yellow gradient follows the N- to C-terminus. (E) α-crystallin domain in (A), showing the seven β-strands forming two antiparallel β-sheets. Color is the same as in (D). (F) crystal structure of the Serine Rich Repeat Protein Binding Region (SRRP-BR) from the gram-positive bacterium *L. reuteri*, which adopts a β-helix fold.

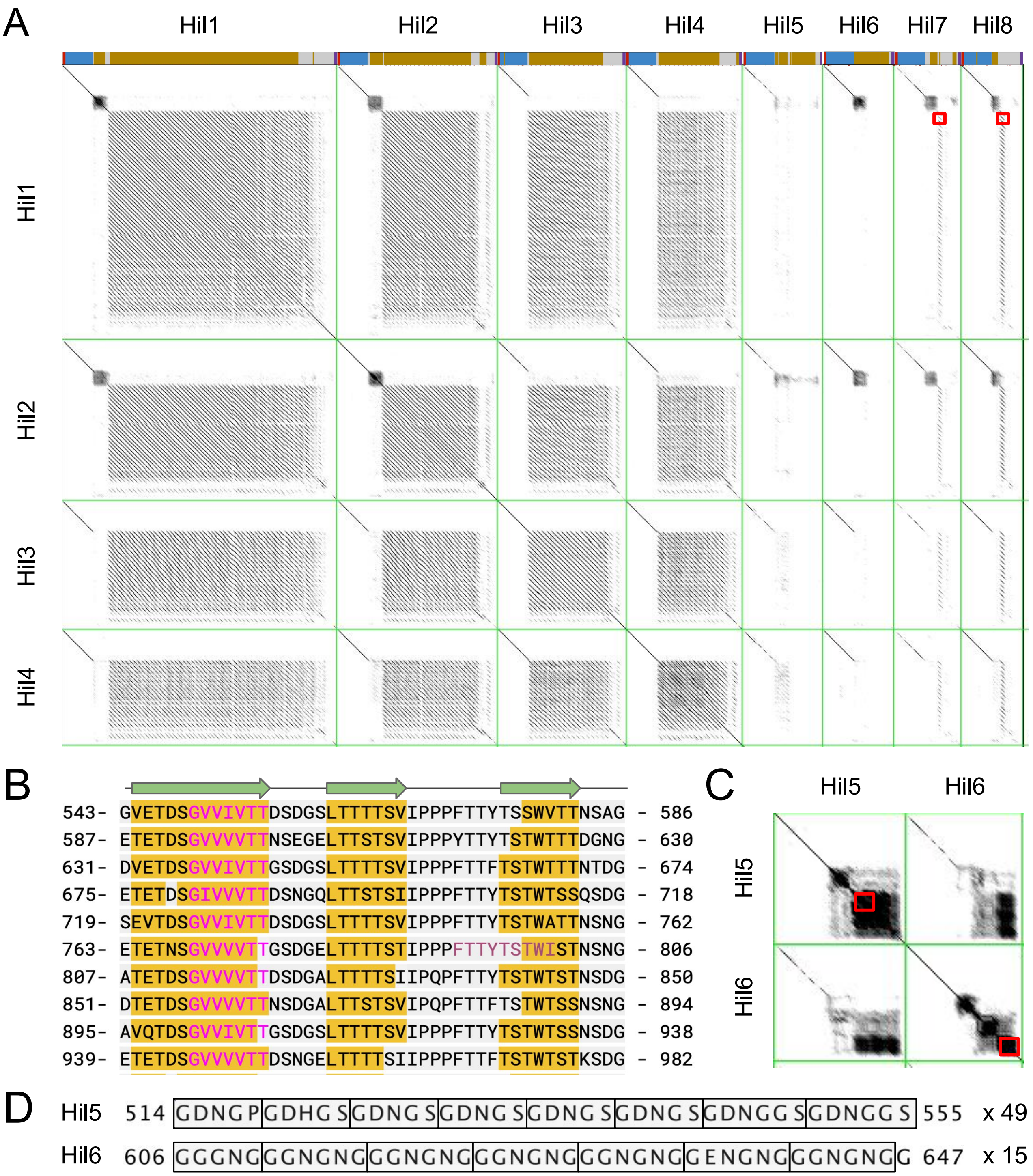
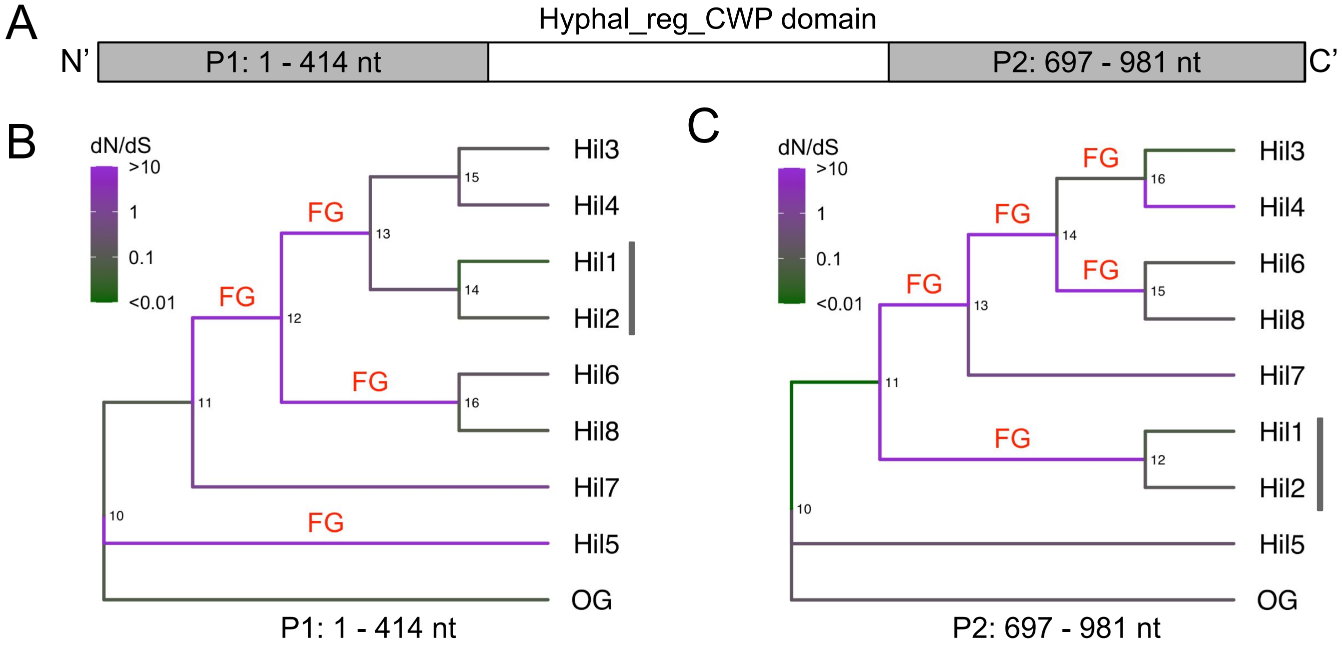


Figure 4. Dotplot shows the tandem repeat structure within and similarity between *C. auris* Hil proteins. (A) Hil1-Hil4 are compared to all eight Hil proteins in *C. auris* including themselves in dotplots (JDotter, Brodie et al 2004) with a sliding window of 50 aa and Grey Map set to 60-245 (min-max). A schematic for each protein is shown above each column (colors same as in Fig. 2). The regions highlighted by the red boxes in row 1 reveal the presence of a single copy of the 44 aa repeat unit in Hil7 and Hil8. (B) Wrapped sequence of aa 543-982 from Hil1 showing the conserved period and sequence of the 44 aa tandem repeat. The magenta and plum fonts indicate motifs predicted by TANGO to have strong (probability > 90%) or moderate (30-90%) β-aggregation potential. The yellow highlighted regions are predicted to form β-strands by PSIPred, with cartoons shown above. (C) Dotplots between Hil5 and Hil6 with the same settings as in (A), showing the low complexity repeats unique to these two. Regions within the two red boxes are shown in (D), with residue numbers shown on both ends. The rectangles delineate individual repeats, with the number of copies for each repeat shown to the right.



D

Tests and models compared	Partition 1-414 nt	Partition 697- 987 nt
a. branch test for variable dN/dS (=ω) between FG=foreground and the rest branches (BG)		
H_0: single dN/dS (ω) **H_1**: two dN/dS (ω_{FG} free)	*P* < 0.001 ω_{FG} = 1.77, ω_{BG} = 0.13	*P* < 0.01 ω_{FG} = 1.53, ω_{BG} = 0.14
H_0: two dN/dS (ω_{FG} == 1) **H_1**: two dN/dS (ω_{FG} free)	Insig.	Insig.
b. branch-site test for positive selection		
H_0: no positive selection **H_1**: positive selection on specific sites in the FG branches	*P* < 0.01 p(ω>1) = 0.53, ω2 = 224.43 *S39, G112*	*P* < 0.01 p(ω>1) = 0.19, ω2 = 9.85 *A234, D287, S304*

Figure 5. Maximum-likelihood-based analyses for selective pressure variation and role of positive selection on the Hyphal_reg_CWP domain in *C. auris*. (A) Schematic showing the putative non-recombining partitions within the Hyphal_reg_CWP domain determined by GARD (see Fig. S8). The two partitions labeled in gray were studied separately. The numbers refer to the columns in the coding sequence alignment. (B, C) Phylogenetic trees were reconstructed for the two partitions and are shown as a cladogram. The vertical bar next to the Hil1/Hil2 pair indicates the difference in topology between the two trees. Branch colors are based on the dN/dS values estimated from a free-ratio model in PAML. "FG" designate foreground branches, whose dN/dS were greater than 10, except for branch 14..16 in (C), which was selected instead of 16..Hil4 because this would require fewer evolutionary changes in selective forces. We also analyzed the scenario with 16..Hil4 as the foreground and the conclusions remained the same with slightly different *P*-values. (D) Summary of the maximum-likelihood-based tests for selective force heterogeneity and for positive selection. "Insig." means *P*-value > 0.05. In the branch-site test, p(ω>1) is the total proportion of sites with dN/dS > 1 on the FG branches, and ω2 their estimated dN/dS. The listed sites were identified as being under positive selection with a posterior probability greater than 0.99 by the Bayes Empirical Bayes (BEB). The one-letter code and number refer to the amino acid in the OG sequence and the alignment column (Fig S7).

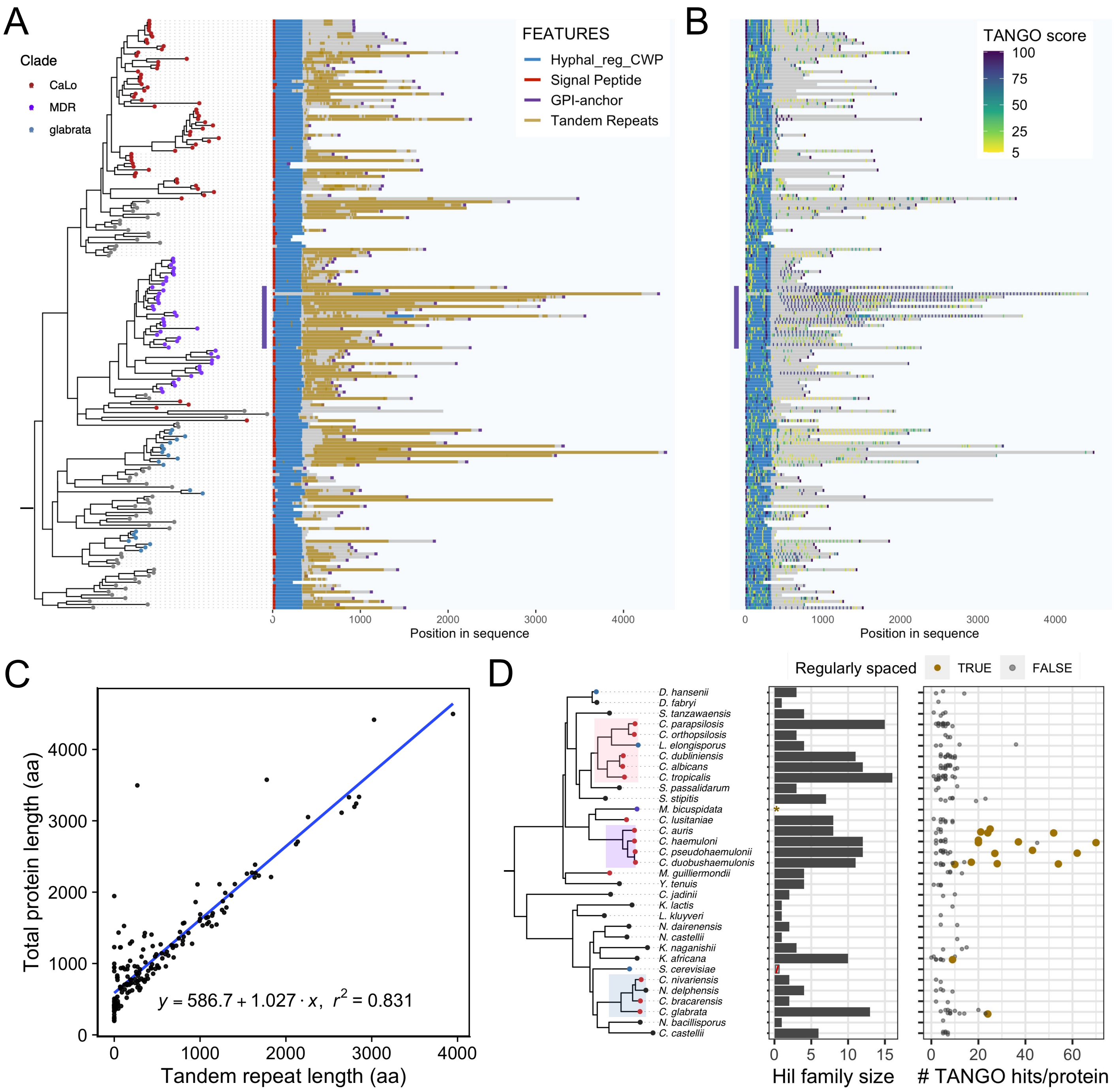


Figure 6. Evolution of protein length and β-aggregation potential in the yeast Hil family. (A) Domain schematic shows that most homologs have a signal peptide at the N-terminus, then the Hyphal_reg_CWP domain and a highly repetitive region central domain, followed by the C-terminal GPI-anchor peptide. Homologs from *M. bicuspidata* were not included because many were annotated as incomplete. They were also excluded from other results in this figure. (B) Distribution of TANGO predicted β-aggregation sequences. The score for each sequence is shown as a color gradient and represents the median of the per-residue probability of aggregation. A vertical bar marks a group of MDR clade sequences that have a large number of β-aggregation prone sequences arranged in regular intervals. (C) X-Y plot showing the relationship between total protein length and tandem repeat sequence length for Hil family homologs. The linear regression line is shown in blue, with coefficients and r^2 values below. (D) The species tree on the left is the same as in Figure 1. The middle panel shows the number of Hil homologs per species. *M. bicuspidata* homologs were excluded; *S. cerevisiae* was included in the species tree but no Hil homolog was identified in it (see text). The right panel shows the number of predicted β-aggregation prone motifs per Hil homolog. Only motifs with a median probability >= 30% were counted. Proteins are colored in gold if they have five or more such motifs and if the Median Absolute Deviation (MAD) of the inter-motif distances is < 5 aa.

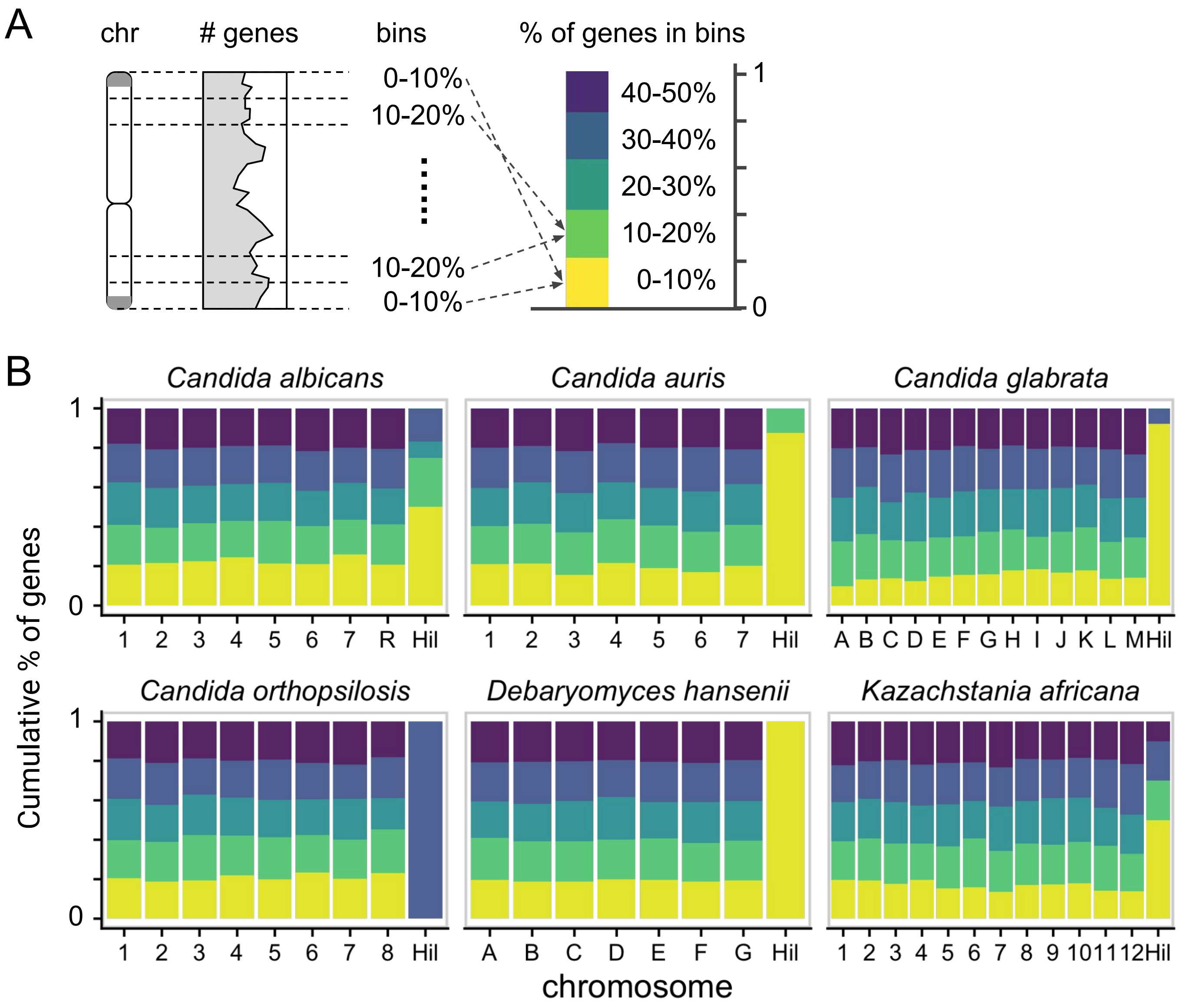
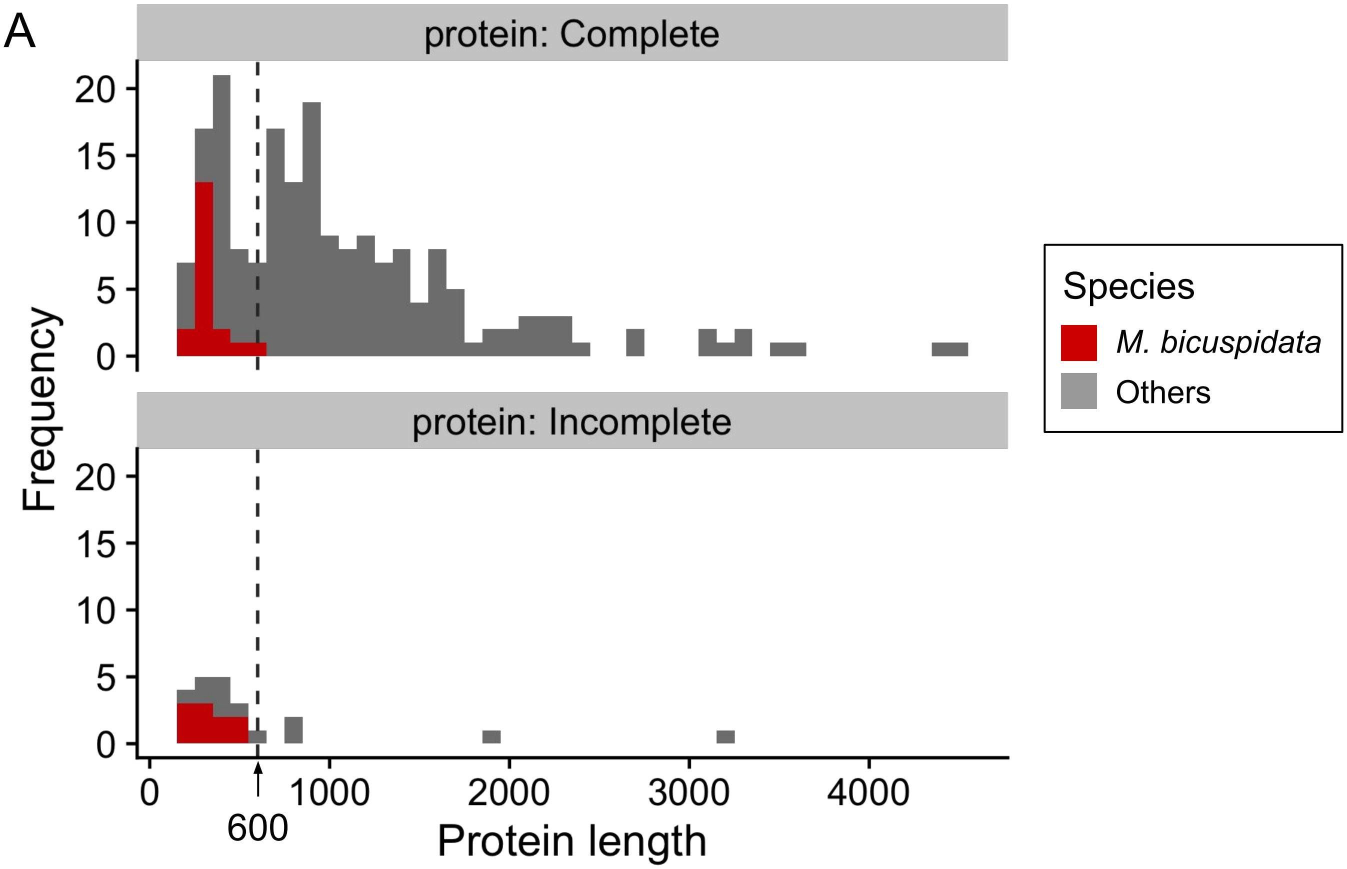


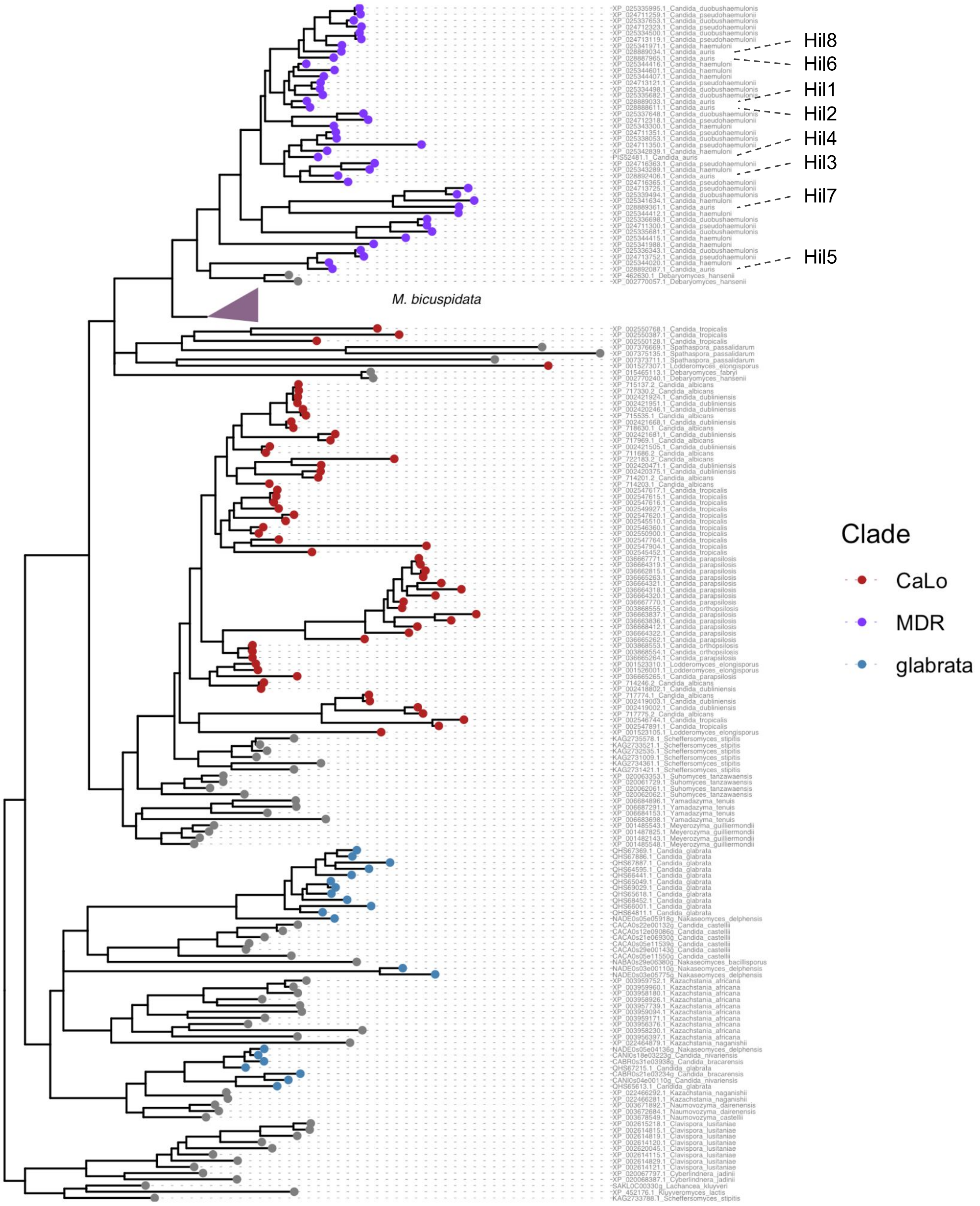
Figure 7. Hil family genes are preferentially located near the chromosome ends. (A) Schematic of the analysis: each chromosome (chr) is folded in half and divided into five equal-length "bins", ordered by their distance to the nearest telomere. The cumulative bar graph on the right summarizes the gene density distribution in the five bins. (B) Folded gene density distribution for six species with a chromosomal level assembly and more than two Hil family genes. The bin colors are as shown in (A). The Hil homologs in each species are plotted as a separate group. A goodness-of-fit test comparing the distribution of the Hil family genes to the genome background yielded a *P*-value of 1.3x10^{-12}.



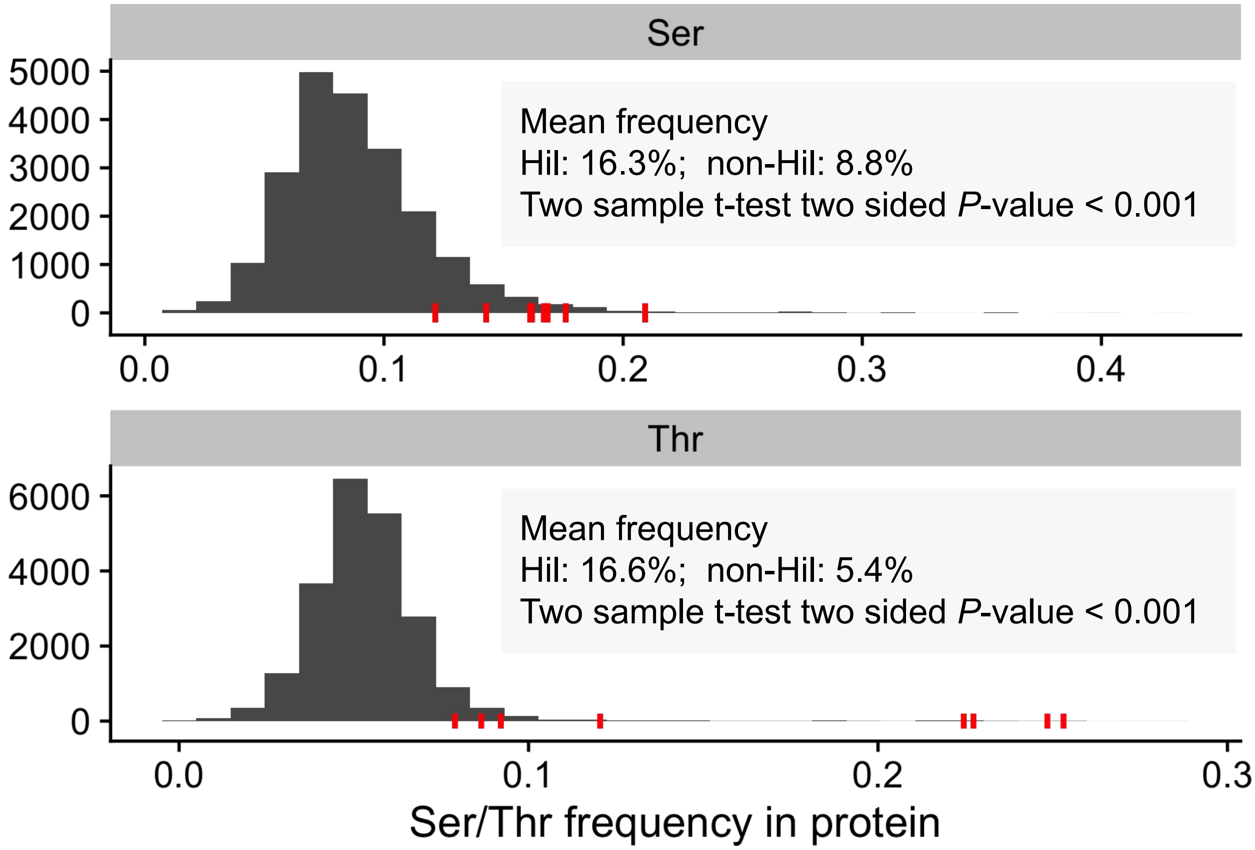
B

length	sp-gpi-	sp-GPI+	SP+gpi-	SP+GPI+	total
>600	3	14	9	112	138
251-600	12	5	27	4	48
0-250	3	0	3	1	7

Supplementary Figure 1. Hil family proteins' length distribution and grouping by signal peptide (SP) and GPI-anchor signal presence. (A) Histogram showing the distribution of protein lengths for Hil family proteins from 32 yeast species. Top: protein sequence records were labeled as complete or "NA"; bottom: proteins labeled as incomplete (no-right, no-left, no-ends). Most of the short sequences (<600 aa, dashed vertical line) came from the species *M. bicuspidata* (red) (B) Summary of the number of Hil family proteins predicted to have a signal peptide (SP+) and GPI-anchor signal (GPI+), grouped by protein length. Proteins labeled as incomplete were excluded from this table.


XP_025335995.1_Candida_duobushaemulonis
XP_024711259.1_Candida_pseudohaemulonii
XP_025337653.1_Candida_duobushaemulonis
XP_024712323.1_Candida_pseudohaemulonii
XP_025334500.1_Candida_duobushaemulonis
XP_024713119.1_Candida_pseudohaemulonii
XP_025341971.1_Candida_haemuloni
XP_028889034.1_Candida_auris
XP_028887965.1_Candida_auris
XP_025344416.1_Candida_haemuloni
XP_025344601.1_Candida_haemuloni
XP_025344407.1_Candida_haemuloni
XP_024713121.1_Candida_pseudohaemulonii
XP_025334498.1_Candida_duobushaemulonis
XP_025335682.1_Candida_duobushaemulonis
XP_028889033.1_Candida_auris
XP_028888611.1_Candida_auris
XP_025337648.1_Candida_duobushaemulonis
XP_024712318.1_Candida_pseudohaemulonii
XP_025343300.1_Candida_haemuloni
XP_024711351.1_Candida_pseudohaemulonii
XP_025338053.1_Candida_duobushaemulonis
XP_024711350.1_Candida_pseudohaemulonii
XP_025342839.1_Candida_haemuloni
PIS52481.1_Candida_auris
XP_024716363.1_Candida_pseudohaemulonii
XP_025343289.1_Candida_haemuloni
XP_028892406.1_Candida_auris
XP_024716365.1_Candida_pseudohaemulonii
XP_024713725.1_Candida_pseudohaemulonii
XP_025339494.1_Candida_duobushaemulonis
XP_025341634.1_Candida_haemuloni
XP_028889361.1_Candida_auris
XP_025344412.1_Candida_haemuloni
XP_025336698.1_Candida_duobushaemulonis
XP_024711300.1_Candida_pseudohaemulonii
XP_025335681.1_Candida_duobushaemulonis
XP_025344415.1_Candida_haemuloni
XP_025341988.1_Candida_haemuloni
XP_025336343.1_Candida_duobushaemulonis
XP_024713752.1_Candida_pseudohaemulonii
XP_025344020.1_Candida_haemuloni
XP_028892087.1_Candida_auris
XP_462630.1_Debaryomyces_hansenii
XP_002770057.1_Debaryomyces_hansenii
Hil8
Hil6
Hil1
Hil2
Hil4
Hil3
Hil7
Hil5
M. bicuspidata
XP_002550768.1_Candida_tropicalis
XP_002550387.1_Candida_tropicalis
XP_002550128.1_Candida_tropicalis
XP_007376669.1_Spathaspora_passalidarum
XP_007375135.1_Spathaspora_passalidarum
XP_007373711.1_Spathaspora_passalidarum
XP_001527307.1_Lodderomyces_elongisporus
XP_015465113.1_Debaryomyces_fabryi
XP_002770240.1_Debaryomyces_hansenii
XP_715137.2_Candida_albicans
XP_717330.2_Candida_albicans
XP_002421924.1_Candida_dubliniensis
XP_002421951.1_Candida_dubliniensis
XP_002420246.1_Candida_dubliniensis
XP_715535.1_Candida_albicans
XP_002421668.1_Candida_dubliniensis
XP_718630.1_Candida_albicans
XP_002421681.1_Candida_dubliniensis
XP_717969.1_Candida_albicans
XP_002421505.1_Candida_dubliniensis
XP_711686.2_Candida_albicans
XP_722183.2_Candida_albicans
XP_002420471.1_Candida_dubliniensis
XP_002420375.1_Candida_dubliniensis
XP_714201.2_Candida_albicans
XP_714203.1_Candida_albicans
XP_002547617.1_Candida_tropicalis
XP_002547615.1_Candida_tropicalis
XP_002547616.1_Candida_tropicalis
XP_002549927.1_Candida_tropicalis
XP_002547620.1_Candida_tropicalis
XP_002545510.1_Candida_tropicalis
XP_002546360.1_Candida_tropicalis
XP_002550900.1_Candida_tropicalis
XP_002547764.1_Candida_tropicalis
XP_002547904.1_Candida_tropicalis
XP_002545452.1_Candida_tropicalis
XP_036667771.1_Candida_parapsilosis
XP_036664319.1_Candida_parapsilosis
XP_036662815.1_Candida_parapsilosis
XP_036665263.1_Candida_parapsilosis
XP_036664321.1_Candida_parapsilosis
XP_036664318.1_Candida_parapsilosis
XP_036664320.1_Candida_parapsilosis
XP_036667770.1_Candida_parapsilosis
XP_003868555.1_Candida_orthopsilosis
XP_036663837.1_Candida_parapsilosis
XP_036663836.1_Candida_parapsilosis
XP_036668412.1_Candida_parapsilosis
XP_036664322.1_Candida_parapsilosis
XP_036665262.1_Candida_parapsilosis
XP_003868553.1_Candida_orthopsilosis
XP_003868554.1_Candida_orthopsilosis
XP_036665264.1_Candida_parapsilosis
XP_001523310.1_Lodderomyces_elongisporus
XP_001526001.1_Lodderomyces_elongisporus
XP_036665265.1_Candida_parapsilosis
XP_714246.2_Candida_albicans
XP_002418802.1_Candida_dubliniensis
XP_717774.1_Candida_albicans
XP_002419003.1_Candida_dubliniensis
XP_002419002.1_Candida_dubliniensis
XP_717775.2_Candida_albicans
XP_002546744.1_Candida_tropicalis
XP_002547891.1_Candida_tropicalis
XP_001523105.1_Lodderomyces_elongisporus
KAG2735578.1_Scheffersomyces_stipitis
KAG2733521.1_Scheffersomyces_stipitis
KAG2732535.1_Scheffersomyces_stipitis
KAG2731009.1_Scheffersomyces_stipitis
KAG2734361.1_Scheffersomyces_stipitis
KAG2731421.1_Scheffersomyces_stipitis
XP_020063353.1_Suhomyces_tanzawaensis
XP_020061729.1_Suhomyces_tanzawaensis
XP_020062061.1_Suhomyces_tanzawaensis
XP_020062062.1_Suhomyces_tanzawaensis
XP_006684896.1_Yamadazyma_tenuis
XP_006687291.1_Yamadazyma_tenuis
XP_006684153.1_Yamadazyma_tenuis
XP_006683698.1_Yamadazyma_tenuis
XP_001485543.1_Meyerozyma_guilliermondii
XP_001487825.1_Meyerozyma_guilliermondii
XP_001482143.1_Meyerozyma_guilliermondii
XP_001485548.1_Meyerozyma_guilliermondii
QHS67369.1_Candida_glabrata
QHS67886.1_Candida_glabrata
QHS67887.1_Candida_glabrata
QHS64595.1_Candida_glabrata
QHS66441.1_Candida_glabrata
QHS65049.1_Candida_glabrata
QHS69029.1_Candida_glabrata
QHS65618.1_Candida_glabrata
QHS68452.1_Candida_glabrata
QHS66001.1_Candida_glabrata
QHS64811.1_Candida_glabrata
NADE0s05e05918g_Nakaseomyces_delphensis
CACA0s22e00132g_Candida_castellii
CACA0s12e09086g_Candida_castellii
CACA0s21e06930g_Candida_castellii
CACA0s05e11539g_Candida_castellii
CACA0s29e00143g_Candida_castellii
CACA0s05e11550g_Candida_castellii
NABA0s29e06380g_Nakaseomyces_bacillisporus
NADE0s03e00110g_Nakaseomyces_delphensis
NADE0s03e05775g_Nakaseomyces_delphensis
XP_003959752.1_Kazachstania_africana
XP_003959960.1_Kazachstania_africana
XP_003958180.1_Kazachstania_africana
XP_003958926.1_Kazachstania_africana
XP_003957739.1_Kazachstania_africana
XP_003959094.1_Kazachstania_africana
XP_003959171.1_Kazachstania_africana
XP_003956376.1_Kazachstania_africana
XP_003958230.1_Kazachstania_africana
XP_003956397.1_Kazachstania_africana
XP_022464879.1_Kazachstania_naganishii
NADE0s05e04136g_Nakaseomyces_delphensis
CANI0s18e03223g_Candida_nivariensis
CABR0s31e03938g_Candida_bracarensis
QHS67215.1_Candida_glabrata
CABR0s21e03234g_Candida_bracarensis
CANI0s04e00110g_Candida_nivariensis
QHS65613.1_Candida_glabrata
XP_022466292.1_Kazachstania_naganishii
XP_022466281.1_Kazachstania_naganishii
XP_003671892.1_Naumovozyma_dairenensis
XP_003672684.1_Naumovozyma_dairenensis
XP_003678549.1_Naumovozyma_castellii
XP_002615218.1_Clavispora_lusitaniae
XP_002614815.1_Clavispora_lusitaniae
XP_002614819.1_Clavispora_lusitaniae
XP_002614120.1_Clavispora_lusitaniae
XP_002620045.1_Clavispora_lusitaniae
XP_002614115.1_Clavispora_lusitaniae
XP_002614829.1_Clavispora_lusitaniae
XP_002614121.1_Clavispora_lusitaniae
XP_020067797.1_Cyberlindnera_jadinii
XP_020068387.1_Cyberlindnera_jadinii
SAKL0C00330g_Lachancea_kluyveri
XP_452176.1_Kluyveromyces_lactis
KAG2733788.1_Scheffersomyces_stipitis
Clade
CaLo
MDR
glabrata

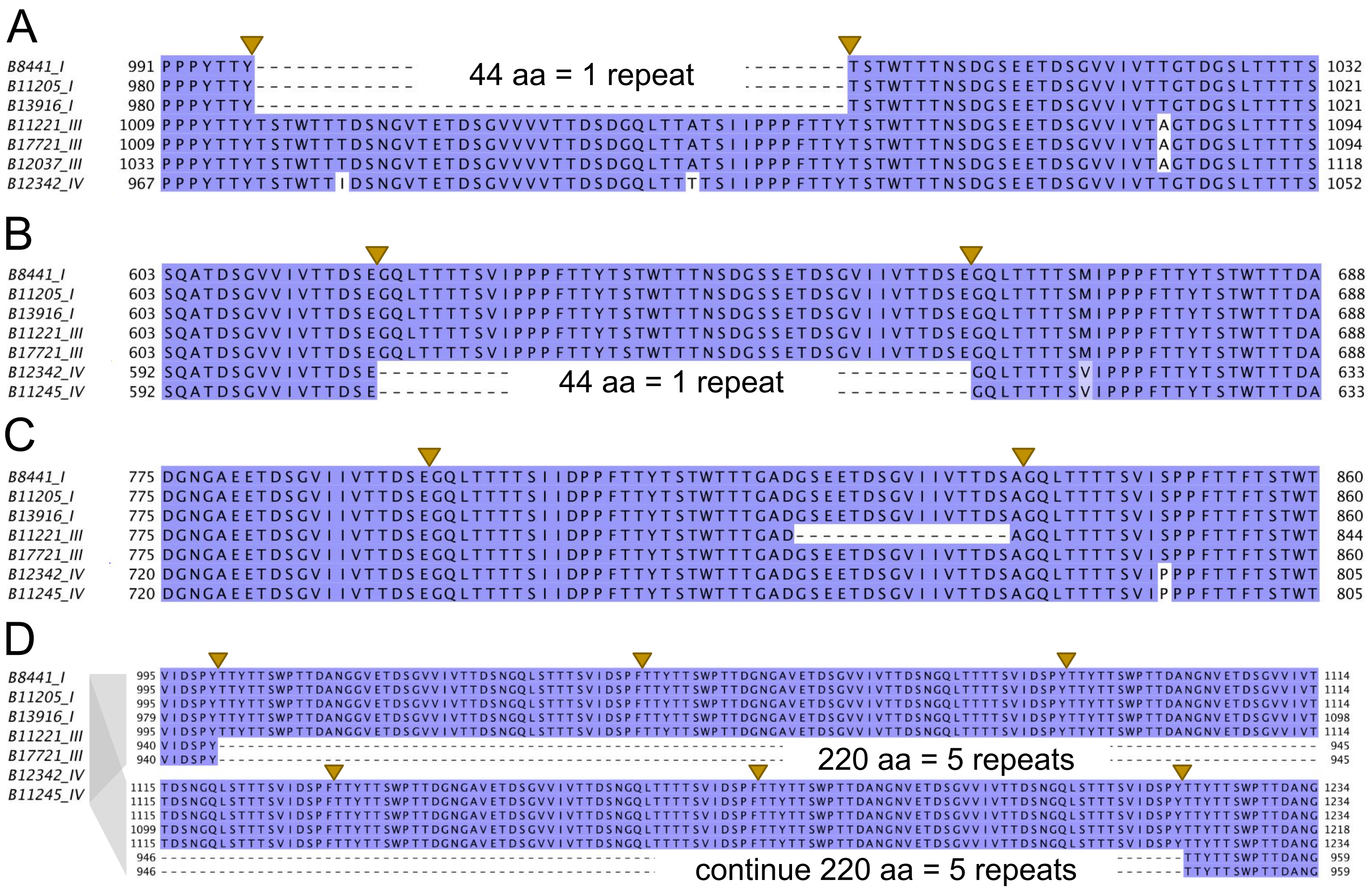
Supplementary Figure 2. Maximum likelihood tree for the Hil family genes. This tree is identical to the one shown in Fig 1C, except that it is shown in a rectangular format with the sequence names in the form of refseqID_species_name. We identified and named the *C. auris* homologs as Hil1-8 to be consistent with the latter figures. Note that all 29 *M. bicuspidata* homologs form a single clade, which is collapsed for the ease of viewing. Their sequence IDs can be found in Table S1.



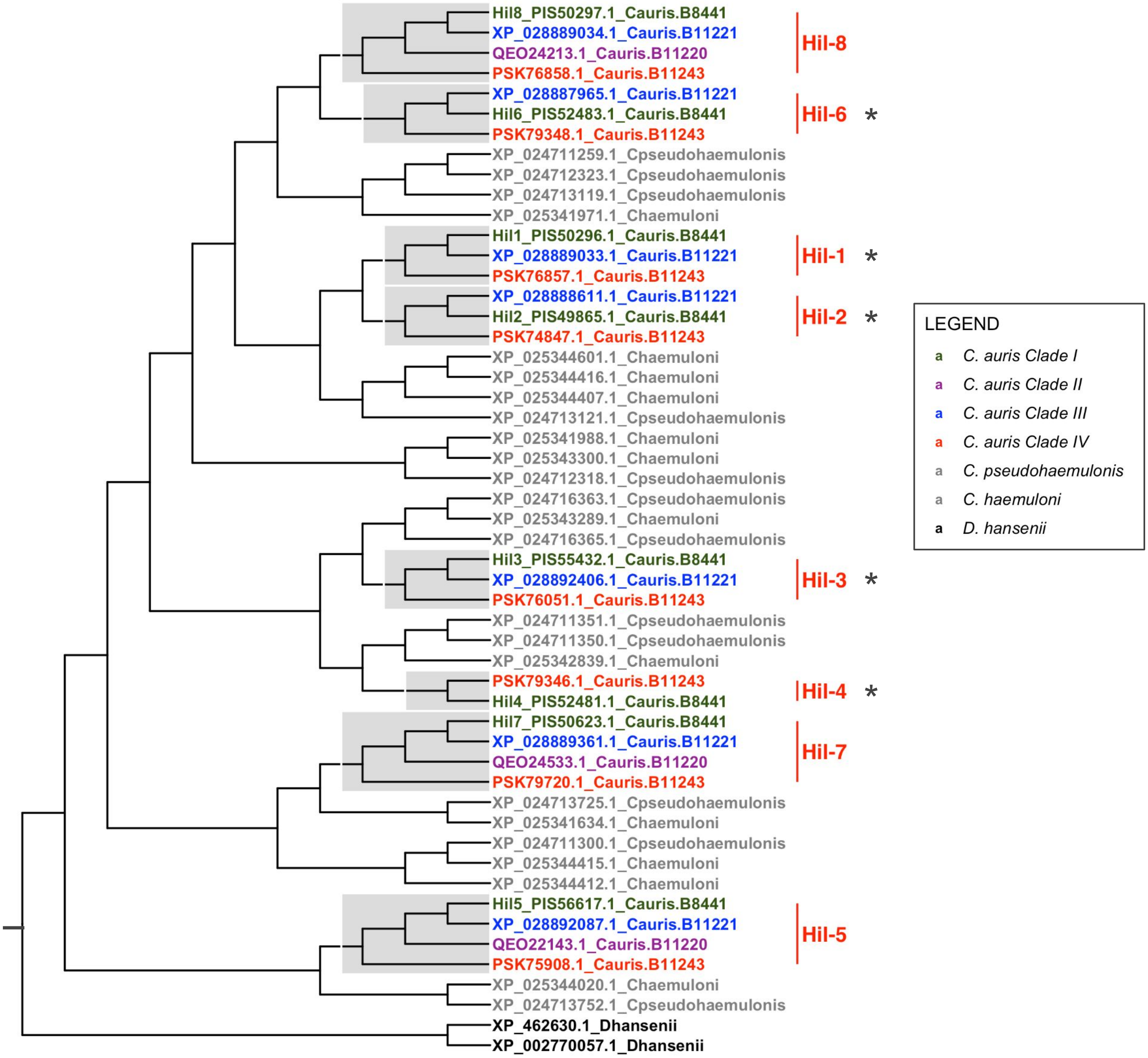
Supplementary Figure 3. Comparison of the Ser/Thr frequencies in *C. auris* Hil family members with all protein-coding genes in *C. auris*. B8441 strain genome is used for this analysis. The frequency of Ser or Thr residues as a percent of the entire protein length is plotted as a histogram for all protein-coding genes. Red ticks indicate the eight Hil genes. A student's t-test was used to assess the significance of the difference in Ser/Thr frequencies between the Hil family proteins vs the rest of the proteome.

```
1  DSGVVIVTTDSDGSLTTTTSVIPPPFTTYTSSWVTTNSAGETET
2  DSGVVVVTTNSEGELTTSTSVIPPPYTTYTSTWTTTDGNGDVET
3  DSGVVIVTTGSDGSLTTTTSVIPPPFTTFTSTWTTTNTDGETET
4  DSGIVVVTTDSNGQLTTSTSIIPPPFTTYTSTWTSSQSDGSEVT
5  DSGVVIVTTDSDGSLTTTTSVIPPPFTTYTSTWATTNSNGETET
6  NSGVVVVTTGSDGELTTTTSTIPPPFTTYTSTWISTNSNGATET
7  DSGVVVVTTDSDGALTTTTSIIPQPFTTYTSTWTSTNSDGDTET
8  DSGVVVVTTNSDGALTTSTSVIPQPFTTFTSTWTSSNSNGAVQT
9  DSGVVIVTTGSDGSLTTTTSVIPPPFTTYTSTWTSSNSDGETET
10 DSGVVVVTTDSNGELTTTTSIIPPPFTTFTSTWTSTKSDGAVET
11 DSGVIIVTTNSEGDLTTTTSIIPPPYTTYTSTWTTTDSNGVTET
12 DSGVVVVTTDSDGQLTTATSIIPPPFTTYTSTWTTTNSDGSEET
13 DSGVVIVTAGTDGSLTTTTSVIPPPFTTYTSTWITTNSNGAVET
14 DSGIIVVTTNSGGSLTTSTSVLPTPFTTYTSTWTTSDGDGNVQT
15 DSGVVIVTTGSDGALSTTTSVIPPPFTTYTSTWISTNSDGETET
16 DSGVVVVTTDSNGALTTTTSIIPPPFTTFTSTWTTTDENGATET
17 DSGVVVVTTGTDGSLTTTTSVIPPPYTTFTSTWTTSNSNGDIET
18 DSGVVIVTTNSDGSLTTTTSVIPPPYTTFTTTWATTNSDGTTET
19 DSGVVIVTTDSEGQLTTTTSVIPPPFTTYTSTWTSNKSDGAVET
20 DSGVVIVTTDSDGALTTTTSIIPQPFTTYTSTWTSTNSNGAIET
21 ESGVVVVTTDSNGALTTSTSVIPLPLTTFTTTWTTTNSAGETET
22 DSGVVVVETNSNGALTTTTSTFPEPFTTFTSTWTTTDDSGAIAT
23 DSGVVIVTTGSDGSLSTTTSVIPPPFTTYTTTWTSTNSNGGVET
24 DSGVVIVTTNSDGALETTTSVIDPPFNTYTSTWTTTDADGAIET
25 DSGVVVVTTGSDGSLTTTTSVIPHPFTTYTSTWVTTGSDGDTET
26 DSGVIVVTTDSDGALTTSTSLLPVPFTTYTSTWTITNSDGSQAT
27 DSGVVIVTTDSEGQLTTTTSVIPPPFTTYTSTWTTTGANGGEET
28 DSGVIIVTTDSDGQLATTTSVIPPPFTTFTSTWTTTNSDGNQAT
29 DSGVVIVTTDSDGQLTTTTSVIPPPFTTYTSTWTTTDGNGAEET
30 DSGVIIVTTDSEGQLTTTTSVIPPPFTTYTSTWTTTGADGSEET
31 DSGVIIVTTDSAGQLTTTTSVIPPPFTTFTSTWTTTDGNGNEGT
32 DSGVIIVTTDSDGALTTTTAVIPPAAGSGTDALSSSINDVPYTTYTSTWTTTDGNGNIET
33 DSGVVIVTTDSQGSLTTTTSIIDSPFTTYTSTWATTDNNGNVET
34 DSGVVIVTTDSNGQLTTTTSVIDSPYTTYTTSWPTTDANGGVET
35 DSGVVIVTTDSDGQLSTTTSVIDSPFTTYTTSWPTTDGNGAVET
36 DSGVVIVTTDSNGQLTTTTSVIDSPYTTYTTSWPTTGADGAVET
37 NSGVVIVTTDSDGQLTTTTSVIDSPYTTYTSIWTTTDSVGNVET
38 DSGVVIVTTDSDGQVTTTTSRFENSPSDLTEYTTTWASTDSDGNIKT
39 DSGVVVVTTDSAGSTTTSTSTFDTPYTTFTSTWTTTNGNGDVKT
40 DSGVIIVTTDSVGQLTTTTSQFDSQQSGLTDYTTTWTTTDRNGNPST
41 ASGVVVVTTDSDGQITSTTSQFSDKSSGLTDYTTTWTTTDTDGSVVT
42 DSGVVIVTTDSAGSLTTTTSVFDTPITTFTSTWTTTNADGSIET
43 DSGVIIVTADSNGQLTTTTSQSDNRPSGLTDYTTTWTTTNTDGAIET
44 DSGVVVVTTDSQGQLTTITSVIESPVTASSGSSDKPSGITEFTTTWTTTDANGIAHT
45 DSGVVIVTTDADGSLTTTTSQIDNVSSGLTEFTSSWTTTLSDGSVET
46 DSGLVIVTTDSNGQLKTTTSQFEDIPSGLSEFTTSWTTTDADGDTRI
47 DSGVVIVTTDSDNRLTTTTSQFASVDPTDFTSYITSWTATNGDGSIET
48 DSGAVIVTTNSDGQLVTTTSVISSSHGAVSTSES
49 DS-NVIVTTDSEGSLTTSTVTLCPQCTHFTSTWTTSNSEGAIET
50 DSGVVVVTTDSVGSLTTYTKDCPEASGELSTFISTYTTTDTDGNIKTT
```

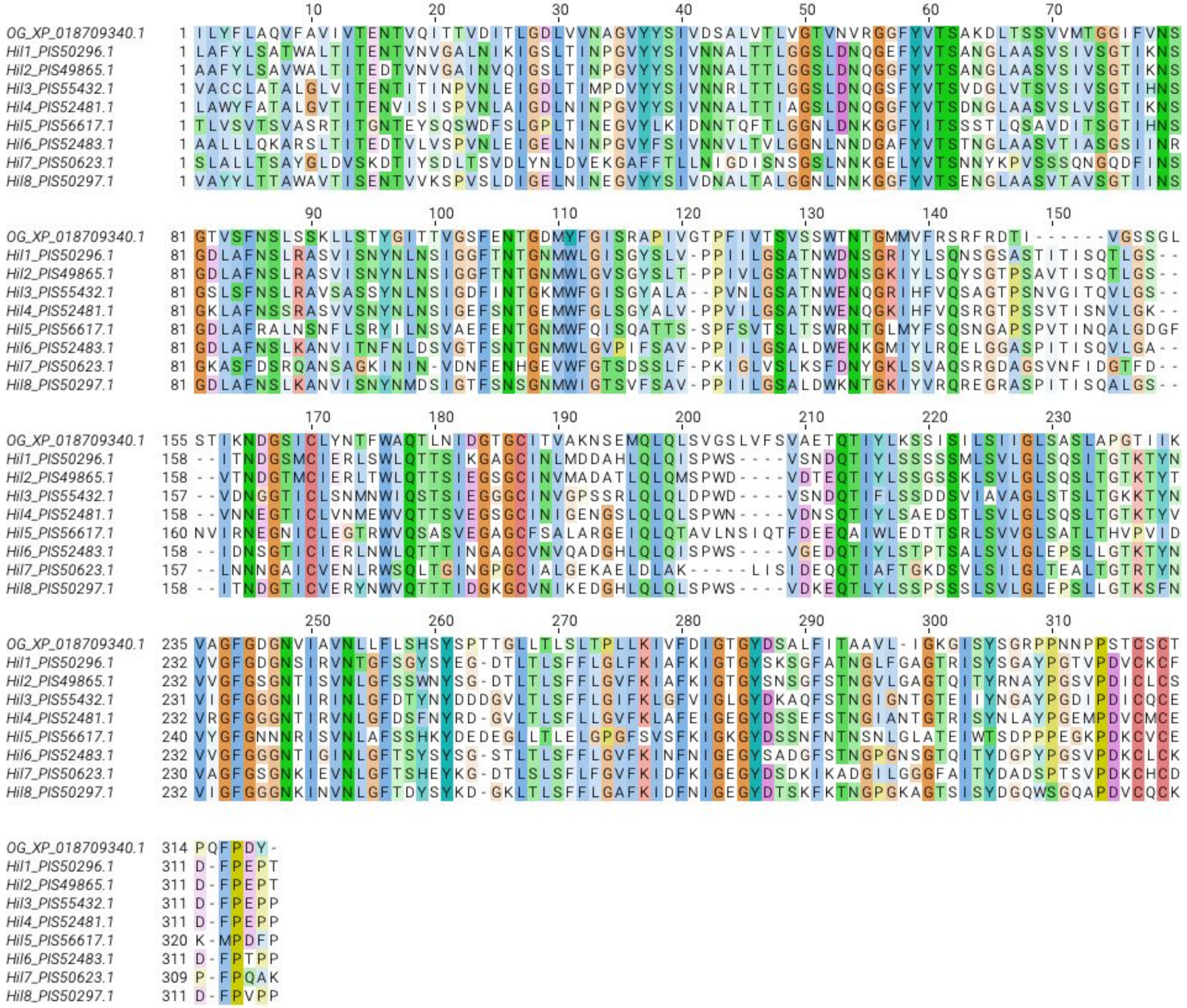
Supplementary Figure 4. Tandem repeats in the *C. auris* Hil1 central domain. The majority of the 50 tandem repeat copies have a conserved 44 aa period. Dark and light orange highlights show sequences predicted by TANGO to have strong (>90%) or moderate (30-90%) β-aggregation potentials.


A
B8441_I 991 PPPYTTY 44 aa = 1 repeat TSTWTTTNSDGSEETDSGVVIVTTGTDGSLTTTTS 1032
B11205_I 980 PPPYTTY TSTWTTTNSDGSEETDSGVVIVTTGTDGSLTTTTS 1021
B13916_I 980 PPPYTTY TSTWTTTNSDGSEETDSGVVIVTTGTDGSLTTTTS 1021
B11221_III 1009 PPPYTTYTSTWTTTDSNGVTETDSGVVVVTTDSDGQLTTATSIIPPPFTTYTSTWTTTNSDGSEETDSGVVIVTAGTDGSLTTTTS 1094
B17721_III 1009 PPPYTTYTSTWTTTDSNGVTETDSGVVVVTTDSDGQLTTATSIIPPPFTTYTSTWTTTNSDGSEETDSGVVIVTAGTDGSLTTTTS 1094
B12037_III 1033 PPPYTTYTSTWTTTDSNGVTETDSGVVVVTTDSDGQLTTATSIIPPPFTTYTSTWTTTNSDGSEETDSGVVIVTAGTDGSLTTTTS 1118
B12342_IV 967 PPPYTTYTSTWTTIDSNGVTETDSGVVVVTTDSDGQLTTTTSIIPPPFTTYTSTWTTTNSDGSEETDSGVVIVTTGTDGSLTTTTS 1052
B
B8441_I 603 SQATDSGVVIVTTDSEGQLTTTTSVIPPPFTTYTSTWTTTNSDGSSETDSGVIIVTTDSEGQLTTTTSMIPPPFTTYTSTWTTTDA 688
B11205_I 603 SQATDSGVVIVTTDSEGQLTTTTSVIPPPFTTYTSTWTTTNSDGSSETDSGVIIVTTDSEGQLTTTTSMIPPPFTTYTSTWTTTDA 688
B13916_I 603 SQATDSGVVIVTTDSEGQLTTTTSVIPPPFTTYTSTWTTTNSDGSSETDSGVIIVTTDSEGQLTTTTSMIPPPFTTYTSTWTTTDA 688
B11221_III 603 SQATDSGVVIVTTDSEGQLTTTTSVIPPPFTTYTSTWTTTNSDGSSETDSGVIIVTTDSEGQLTTTTSMIPPPFTTYTSTWTTTDA 688
B17721_III 603 SQATDSGVVIVTTDSEGQLTTTTSVIPPPFTTYTSTWTTTNSDGSSETDSGVIIVTTDSEGQLTTTTSMIPPPFTTYTSTWTTTDA 688
B12342_IV 592 SQATDSGVVIVTTDSE 44 aa = 1 repeat GQLTTTTSVIPPPFTTYTSTWTTTDA 633
B11245_IV 592 SQATDSGVVIVTTDSE GQLTTTTSVIPPPFTTYTSTWTTTDA 633
C
B8441_I 775 DGNGAEETDSGVIIVTTDSEGQLTTTTSIIDPPFTTYTSTWTTTGADGSEETDSGVIIVTTDSAGQLTTTTSVISPPFTTFTSTWT 860
B11205_I 775 DGNGAEETDSGVIIVTTDSEGQLTTTTSIIDPPFTTYTSTWTTTGADGSEETDSGVIIVTTDSAGQLTTTTSVISPPFTTFTSTWT 860
B13916_I 775 DGNGAEETDSGVIIVTTDSEGQLTTTTSIIDPPFTTYTSTWTTTGADGSEETDSGVIIVTTDSAGQLTTTTSVISPPFTTFTSTWT 860
B11221_III 775 DGNGAEETDSGVIIVTTDSEGQLTTTTSIIDPPFTTYTSTWTTTGAD AGQLTTTTSVISPPFTTFTSTWT 844
B17721_III 775 DGNGAEETDSGVIIVTTDSEGQLTTTTSIIDPPFTTYTSTWTTTGADGSEETDSGVIIVTTDSAGQLTTTTSVISPPFTTFTSTWT 860
B12342_IV 720 DGNGAEETDSGVIIVTTDSEGQLTTTTSIIDPPFTTYTSTWTTTGADGSEETDSGVIIVTTDSAGQLTTTTSVIPPPFTTFTSTWT 805
B11245_IV 720 DGNGAEETDSGVIIVTTDSEGQLTTTTSIIDPPFTTYTSTWTTTGADGSEETDSGVIIVTTDSAGQLTTTTSVIPPPFTTFTSTWT 805
D
B8441_I
B11205_I
B13916_I
B11221_III
B17721_III
B12342_IV
B11245_IV
995 VIDSPYTTYTTSWPTTDANGGVETDSGVVIVTTDSNGQLSTTTSVIDSPFTTYTTSWPTTDGNGAVETDSGVVIVTTDSNGQLTTTTSVIDSPYTTYTTSWPTTDANGNVETDSGVVIVT 1114
995 VIDSPYTTYTTSWPTTDANGGVETDSGVVIVTTDSNGQLSTTTSVIDSPFTTYTTSWPTTDGNGAVETDSGVVIVTTDSNGQLTTTTSVIDSPYTTYTTSWPTTDANGNVETDSGVVIVT 1114
995 VIDSPYTTYTTSWPTTDANGGVETDSGVVIVTTDSNGQLSTTTSVIDSPFTTYTTSWPTTDGNGAVETDSGVVIVTTDSNGQLTTTTSVIDSPYTTYTTSWPTTDANGNVETDSGVVIVT 1114
979 VIDSPYTTYTTSWPTTDANGGVETDSGVVIVTTDSNGQLSTTTSVIDSPFTTYTTSWPTTDGNGAVETDSGVVIVTTDSNGQLTTTTSVIDSPYTTYTTSWPTTDANGNVETDSGVVIVT 1098
995 VIDSPYTTYTTSWPTTDANGGVETDSGVVIVTTDSNGQLSTTTSVIDSPFTTYTTSWPTTDGNGAVETDSGVVIVTTDSNGQLTTTTSVIDSPYTTYTTSWPTTDANGNVETDSGVVIVT 1114
940 VIDSPY 945
940 VIDSPY 945
220 aa = 5 repeats
1115 TDSNGQLSTTTSVIDSPFTTYTTSWPTTDGNGAVETDSGVVIVTTDSNGQLTTTTSVIDSPFTTYTTSWPTTDANGNVETDSGVVIVTTDSNGQLSTTTSVIDSPYTTYTTSWPTTDANG 1234
1115 TDSNGQLSTTTSVIDSPFTTYTTSWPTTDGNGAVETDSGVVIVTTDSNGQLTTTTSVIDSPFTTYTTSWPTTDANGNVETDSGVVIVTTDSNGQLSTTTSVIDSPYTTYTTSWPTTDANG 1234
1115 TDSNGQLSTTTSVIDSPFTTYTTSWPTTDGNGAVETDSGVVIVTTDSNGQLTTTTSVIDSPFTTYTTSWPTTDANGNVETDSGVVIVTTDSNGQLSTTTSVIDSPYTTYTTSWPTTDANG 1234
1099 TDSNGQLSTTTSVIDSPFTTYTTSWPTTDGNGAVETDSGVVIVTTDSNGQLTTTTSVIDSPFTTYTTSWPTTDANGNVETDSGVVIVTTDSNGQLSTTTSVIDSPYTTYTTSWPTTDANG 1218
1115 TDSNGQLSTTTSVIDSPFTTYTTSWPTTDGNGAVETDSGVVIVTTDSNGQLTTTTSVIDSPFTTYTTSWPTTDANGNVETDSGVVIVTTDSNGQLSTTTSVIDSPYTTYTTSWPTTDANG 1234
946 TTYTTSWPTTDANG 959
946 TTYTTSWPTTDANG 959
continue 220 aa = 5 repeats

Supplementary Figure 5. Examples of tandem repeat copy number variation in Hil1-Hil4 among the *C. auris* strains. (A) A 44 aa indel in Hil1 removes exactly one repeat in all three Clade I strain orthologs. (B) A similar indel polymorphism of exactly one repeat length in Hil2 affecting the Clade IV strains. (C) An indel polymorphism in Hil2 that affects one Clade III strain and spans 16 aa, not a full repeat, but includes a predicted strong β-aggregation prone sequence “GVIIVTT”. (D) An indel polymorphism in Hil2 that spans 220 aa or five full repeats affecting the Clade IV strains. Similar patterns were observed in Hil3 and Hil4.

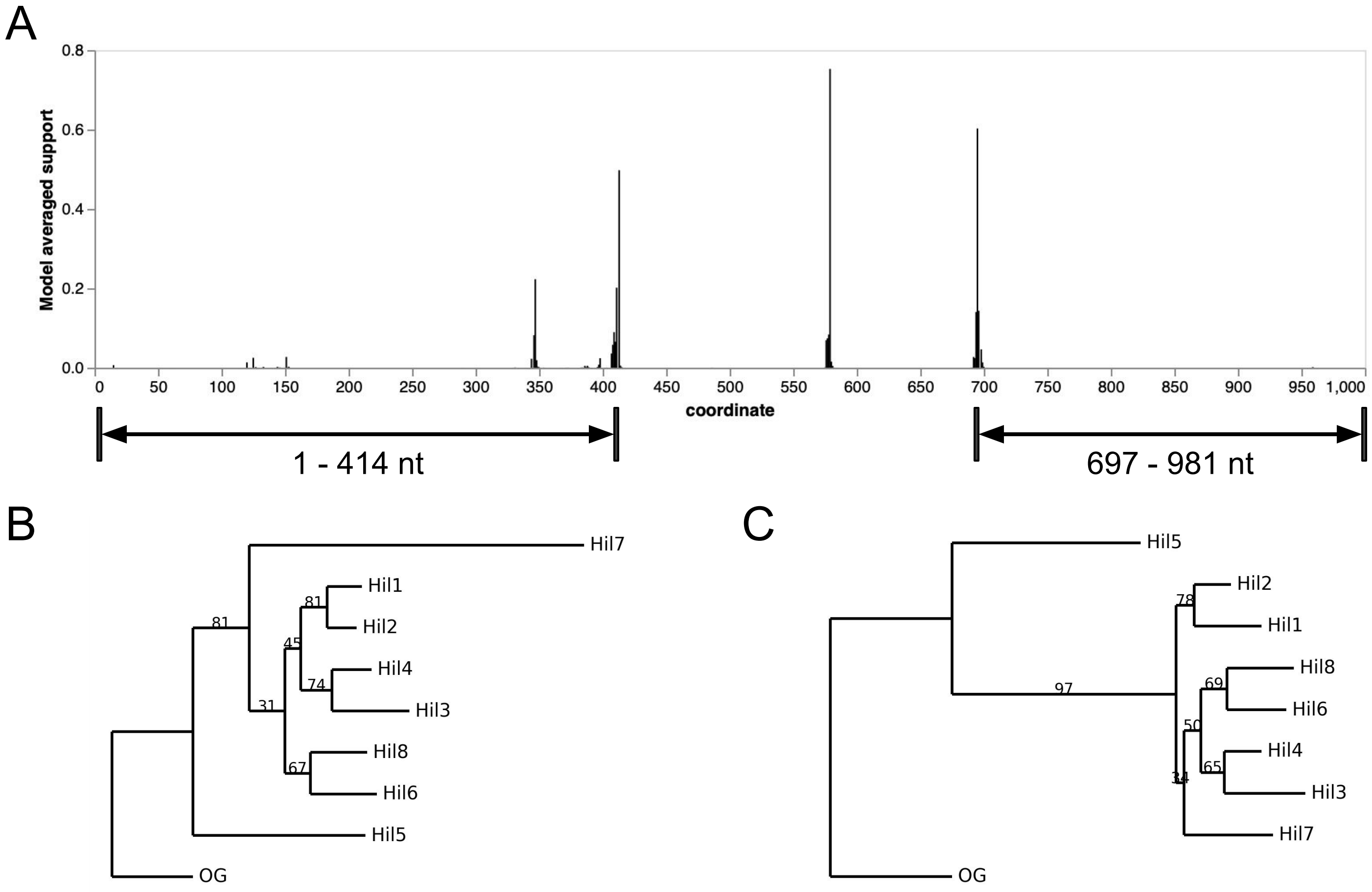

Hil8_PIS50297.1_Cauris.B8441
XP_028889034.1_Cauris.B11221
QEO24213.1_Cauris.B11220
PSK76858.1_Cauris.B11243
Hil-8
XP_028887965.1_Cauris.B11221
Hil6_PIS52483.1_Cauris.B8441
PSK79348.1_Cauris.B11243
Hil-6 *
XP_024711259.1_Cpseudohaemulonis
XP_024712323.1_Cpseudohaemulonis
XP_024713119.1_Cpseudohaemulonis
XP_025341971.1_Chaemuloni
Hil1_PIS50296.1_Cauris.B8441
XP_028889033.1_Cauris.B11221
PSK76857.1_Cauris.B11243
Hil-1 *
XP_028888611.1_Cauris.B11221
Hil2_PIS49865.1_Cauris.B8441
PSK74847.1_Cauris.B11243
Hil-2 *
XP_025344601.1_Chaemuloni
XP_025344416.1_Chaemuloni
XP_025344407.1_Chaemuloni
XP_024713121.1_Cpseudohaemulonis
XP_025341988.1_Chaemuloni
XP_025343300.1_Chaemuloni
XP_024712318.1_Cpseudohaemulonis
XP_024716363.1_Cpseudohaemulonis
XP_025343289.1_Chaemuloni
XP_024716365.1_Cpseudohaemulonis
Hil3_PIS55432.1_Cauris.B8441
XP_028892406.1_Cauris.B11221
PSK76051.1_Cauris.B11243
Hil-3 *
XP_024711351.1_Cpseudohaemulonis
XP_024711350.1_Cpseudohaemulonis
XP_025342839.1_Chaemuloni
PSK79346.1_Cauris.B11243
Hil4_PIS52481.1_Cauris.B8441
Hil-4 *
Hil7_PIS50623.1_Cauris.B8441
XP_028889361.1_Cauris.B11221
QEO24533.1_Cauris.B11220
PSK79720.1_Cauris.B11243
Hil-7
XP_024713725.1_Cpseudohaemulonis
XP_025341634.1_Chaemuloni
XP_024711300.1_Cpseudohaemulonis
XP_025344415.1_Chaemuloni
XP_025344412.1_Chaemuloni
Hil5_PIS56617.1_Cauris.B8441
XP_028892087.1_Cauris.B11221
QEO22143.1_Cauris.B11220
PSK75908.1_Cauris.B11243
Hil-5
XP_025344020.1_Chaemuloni
XP_024713752.1_Cpseudohaemulonis
XP_462630.1_Dhansenii
XP_002770057.1_Dhansenii
LEGEND
a C. auris Clade I
a C. auris Clade II
a C. auris Clade III
a C. auris Clade IV
a C. pseudohaemulonis
a C. haemuloni
a D. hansenii

Supplementary Figure 6. Reconciled Hil family gene tree based on the Hyphal_reg_CWP domain alignment in the four clades of *C. auris* strains and two closely related species. The tree is rooted by the two homologs from the outgroup *D. hansenii*. The gene tree was corrected with the species/strain tree based on (Muñoz et al 2018) using GeneRax (v2.0.4). Hil genes lost in *C. auris* Clade II strains are labeled with an asterisk next to the Hil1-8 group labels.

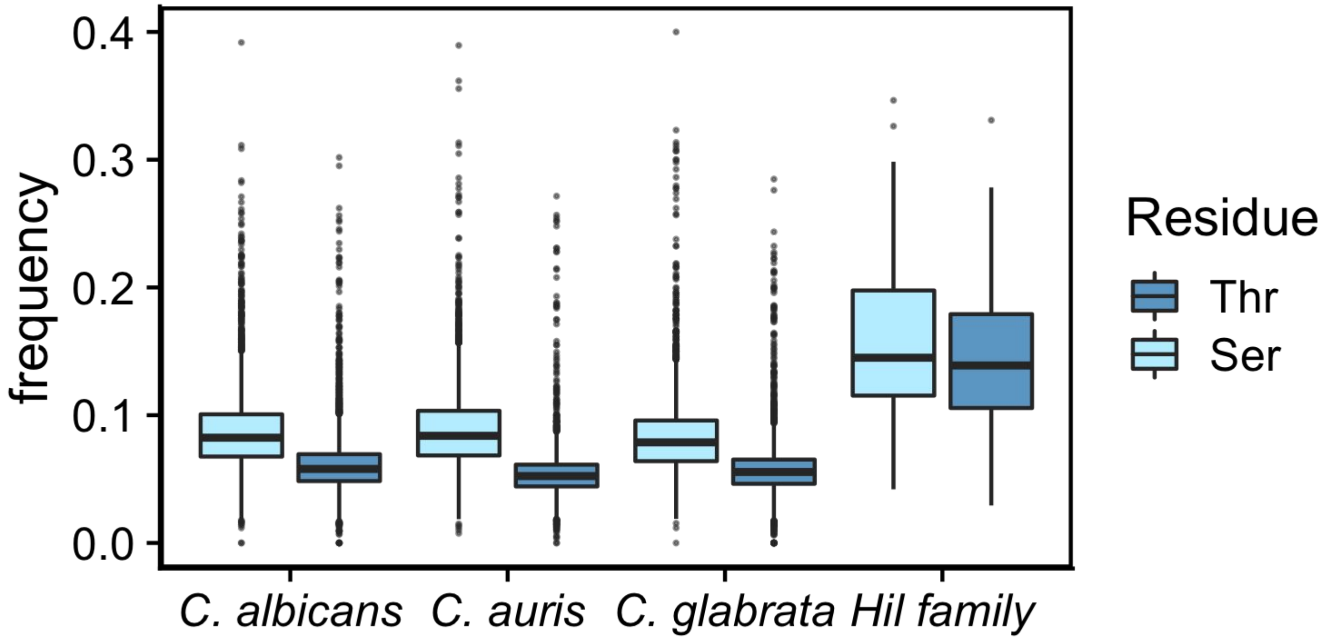


```
                               10        20        30        40        50        60        70
OG_XP_018709340.1   1 ILYFLAQVFAVIVTENTVQITTVDITLGDLVVNAGVYYSIVDSALVTLVGTVNVRGGFYVTSAKDLTSSVVMTGGIFVNS
Hil1_PIS50296.1     1 LAFYLSATWALTITENTVNVGALNIKIGSLTINPGVYYSIVNNALTTLGGSLDNQGEFYVTSANGLAASVSIVSGTIKNS
Hil2_PIS49865.1     1 AAFYLSAVWALTITEDTVNVGAINVQIGSLTINPGVYYSIVNNALTTLGGSLDNQGGFYVTSANGLAASVSIVSGTIKNS
Hil3_PIS55432.1     1 VACCLATALGLVITENTITINPVNLEIGDLTIMPDVYYSIVNNRLTTLGGSLDNQGSFYVTSVDGLVTSVSIVSGTIHNS
Hil4_PIS52481.1     1 LAWYFATALGVTITENVISISPVNLAIGDLNINPGVYYSIVNNALTTIAGSLDNQGGFYVTSDNGLAASVSLVSGTIKNS
Hil5_PIS56617.1     1 TLVSVTSVASRTITGNTEYSQSWDFSLGPLTINEGVYLKIDNNTQFTLGGNLDNKGGFYITSSSTLQSAVDITSGTIHNS
Hil6_PIS52483.1     1 AALLLQKARSLTITEDTVLVSPVNLEIGELNINPGVYFSIVNNVLTVLGGNLNNDGAFYVTSTNGLAASVTIASGSIINR
Hil7_PIS50623.1     1 SLALLTSAYGLDVSKDTIYSDLTSVDLYNLDVEKGAFFTLLNIGDISNSGSLNNKGELYVTSNNYKPVSSSQNGQDFINS
Hil8_PIS50297.1     1 VAYYLTTAWAVTISENTVVKSPVSLDIGELNINEGVYYSIVDNALTALGGNLNNKGGFYVTSENGLAASVTAVSGTIINS

                               90        100       110       120       130       140       150
OG_XP_018709340.1  81 GTVSFNSLSSKLLSTYGITTVGSFENTGDMYFGISRAPIVGTPFIVTSVSSWTNTGMMVFRSRFRDTI------VGSSGL
Hil1_PIS50296.1    81 GDLAFNSLRASVISNYNLNSIGGFTNTGNMWLGISGYSLV-PPIILGSATNWDNSGRIYLSQNSGSASTITISQTLGS--
Hil2_PIS49865.1    81 GDLAFNSLRASVISNYNLNSIGGFTNTGNMWLGVSGYSLT-PPIVLGSATNWDNSGKIYLSQYSGTPSAVTISQTLGS--
Hil3_PIS55432.1    81 GSLSFNSLRAVSASSYNLNSIGDFINTGKMWFGISGYALA--PVNLGSATNWENQGRIHFVQSAGTPSNVGITQVLGS--
Hil4_PIS52481.1    81 GKLAFNSSRASVVSNYNLNSIGEFSNTGEMWFGLSGYALV-PPVILGSATNWENQGKIHFVQSRGTPSSVTISNVLGK--
Hil5_PIS56617.1    81 GDLAFRALNSNFLSRYILNSVAEFENTGNMWFQISQATTS-SPFSVTSLTSWRNTGLMYFSQSNGAPSPVTINQALGDGF
Hil6_PIS52483.1    81 GDLAFNSLKANVITNFNLDSVGTFSNTGNMWLGVPIFSAV-PPIILGSALDWENKGMIYLRQELGGASPITISQVLGA--
Hil7_PIS50623.1    81 GKASFDSRQANSAGKININ-VDNFENHGEVWFGTSDSSLF-PKIGLVSLKSFDNYGKLSVAQSRGDAGSVNFIDGTFD--
Hil8_PIS50297.1    81 GDLAFNSLKANVISNYNMDSIGTFSNSGNMWIGTSVFSAV-PPIILGSALDWKNTGKIYVRQREGRASPITISQALGS--

                               170       180       190       200       210       220       230
OG_XP_018709340.1 155 STIKNDGSICLYNTFWAQTLNIDGTGCITVAKNSEMQLQLSVGSLVFSVAETQTIYLKSSISILSIIGLSASLAPGTIIK
Hil1_PIS50296.1   158 --ITNDGSMCIERLSWLQTTSIKGAGCINLMDDAHLQLQISPWS----VSNDQTIYLSSSSSMLSVLGLSQSITGTKTYN
Hil2_PIS49865.1   158 --VTNDGTMCIERLTWLQTTSIEGSGCINVMADATLQLQMSPWD----VDTEQTIYLSSGSSKLSVLGLSQSLTGTKTYT
Hil3_PIS55432.1   157 --VDNGGTICLSNMNWIQSTSIEGGGCINVGPSSRLQLQLDPWD----VSNDQTIFLSSDDSVIAVAGLSTSLTGKKTYN
Hil4_PIS52481.1   158 --VNNEGTICLVNMEWVQTTSVEGSGCINIGENGSLQLQLSPWN----VDNSQTIYLSAEDSTLSVLGLSQSLTGTKTYV
Hil5_PIS56617.1   160 NVIRNEGNICLEGTRWVQSASVEGAGCFSALARGEIQLQTAVLNSIQTFDEEQAIWLEDTTSRLSVVGLSATLTHVPVID
Hil6_PIS52483.1   158 --IDNSGTICIERLNWLQTTTINGAGCVNVQADGHLQLQISPWS----VGEDQTIYLSTPTSALSVLGLEPSLLGTKTYN
Hil7_PIS50623.1   157 --LNNNGAICVENLRWSQLTGINGPGCIALGEKAELDLAK-----LISIDEQQTIAFTGKDSVLSILGLTEALTGTRTYN
Hil8_PIS50297.1   158 --ITNDGTICVERYNWVQTTTIDGKGCVNIKEDGHLQLQLSPWS----VDKEQTLYLSSPSSSLSVLGLEPSLLGTKSFN

                               250       260       270       280       290       300       310
OG_XP_018709340.1 235 VAGFGDGNVIAVNLLFLSHSYSPTTGLLTLSLTPLLKIVFDIGTGYDSALFITAAVL-IGKGISYSGRPPNNPPSTCSCT
Hil1_PIS50296.1   232 VVGFGDGNSIRVNTGFSGYSYEG-DTLTLSFFLGLFKIAFKIGTGYSKSGFATNGLFGAGTRISYSGAYPGTVPDVCKCF
Hil2_PIS49865.1   232 VVGFGSGNTISVNLGFSSWNYSG-DTLTLSFFLGVFKIAFKIGTGYSNSGFSTNGVLGAGTQITYRNAYPGSVPDICLCS
Hil3_PIS55432.1   231 VIGFGGGNIIRINLGFDTYNYDDDGVLTLSFFLGIFKLGFVIGLGYDKAQFSTNGIGNTGTEIIYNGAYPGDIPDICQCE
Hil4_PIS52481.1   232 VRGFGGGNTIRVNLGFDSFNYRD-GVLTLSFLLGVFKLAFEIGEGYDSSEFSTNGIANTGTRISYNLAYPGEMPDVCMCE
Hil5_PIS56617.1   240 VYGFGNNNRISVNLAFSSHKYDEDEGLLTLELGPGFSVSFKIGKGYDSSNFNTNSNLGLATEIWTSDPPPEGKPDKCVCE
Hil6_PIS52483.1   232 VVGFGGGNTIGINLGFTSYSYSG-STLTLSFFLGVFKINFNIGEGYSADGFSTNGPGNSGTQITYDGPYPGSVPDKCLCK
Hil7_PIS50623.1   230 VAGFGSGNKIEVNLGFTSHEYKG-DTLSLSFLFGVFKIDFKIGEGYDSDKIKADGILGGGFAITYDADSPTSVPDKCHCD
Hil8_PIS50297.1   232 VIGFGGGNKINVNLGFTDYSYKD-GKLTLSFFLGAFKIDFNIGEGYDTSKFKTNGPGKAGTSISYDGQWSGQAPDVCQCK

OG_XP_018709340.1 314 PQFPDY-
Hil1_PIS50296.1   311 D-FPEPT
Hil2_PIS49865.1   311 D-FPEPT
Hil3_PIS55432.1   311 D-FPEPP
Hil4_PIS52481.1   311 D-FPEPP
Hil5_PIS56617.1   320 K-MPDFP
Hil6_PIS52483.1   311 D-FPTPP
Hil7_PIS50623.1   309 P-FPQAK
Hil8_PIS50297.1   311 D-FPVPP
```

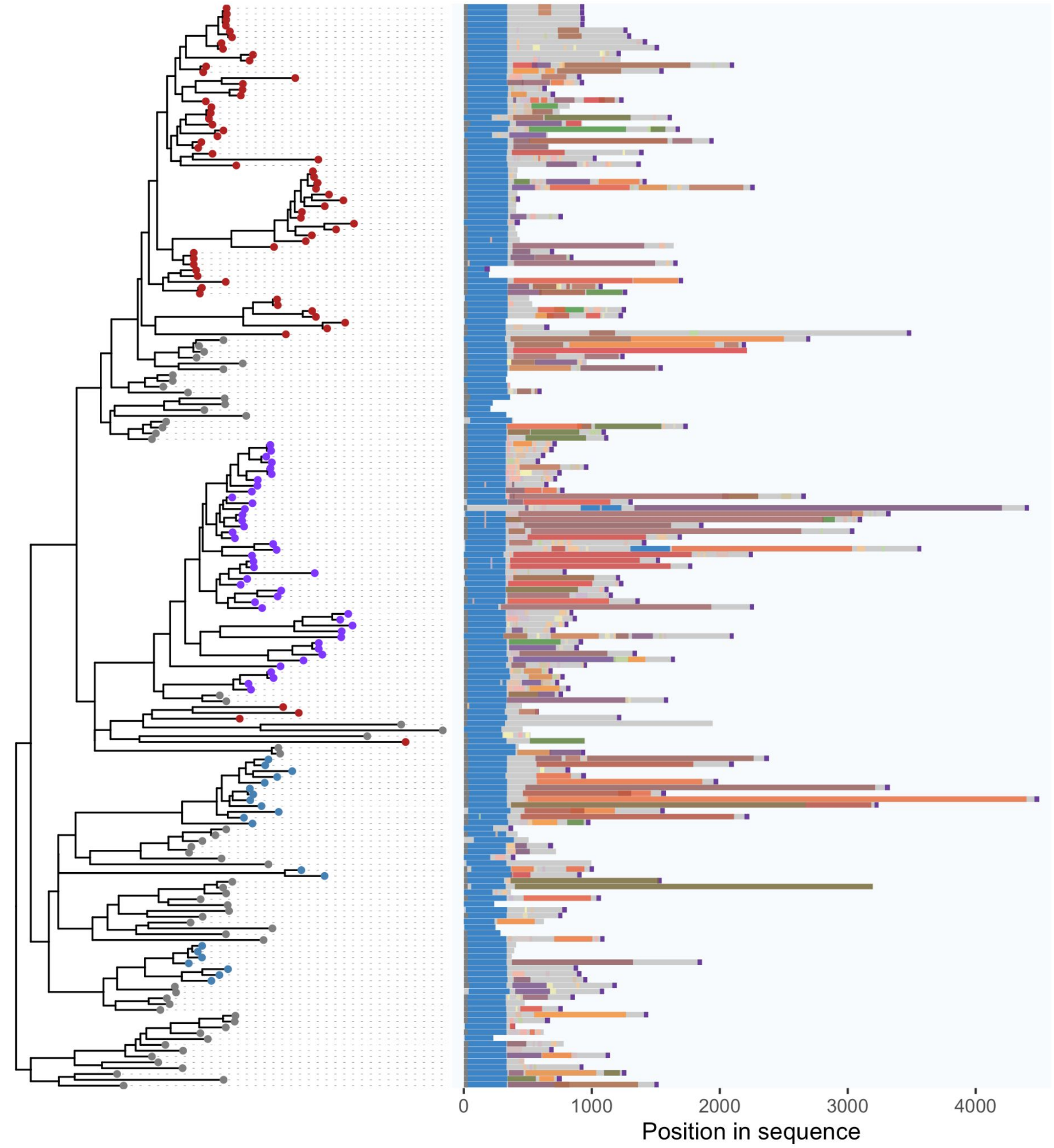
Supplementary Figure 7. Multiple sequence alignment of *C. auris* Hil1-8 Hyphal_reg_CWP domain. The domain regions were identified using HMMScan against the Pfam-A database. The sequences were aligned with clustalo v1.2.4 and the result visualized in Jalview v2.11.1.4 with the ClustalW color scheme. OG = outgroup from *M. bicuspidata*



Supplementary Figure 8. Detecting intra-domain recombination and identifying non-recombining partitions in the Hyphal_reg_CWP domain using GARD. (A) Model averaged support for breakpoint locations along the Hyphal_reg_CWP domain alignment for the eight Hil proteins in *C. auris* and an outgroup sequence from *M. bicuspidata* (protein ID: XP_018709340.1) to root the gene tree. Based on the GARD output, we chose the N- and C-terminal partitions for downstream analyses, i.e., coordinates 1-414 nt and 697-981 nt. (B) A maximum likelihood tree for partition 1-414 was constructed using RAxML-NG v1.1.0. Branch length is proportional to the amount of sequence divergence. OG stands for outgroup. Bootstrap support for internal splits are shown as a percentage and are based on 1000 replicates or until bootstrapping converges. (C) tree for 697-981nt, same format as in (B)



Supplementary Figure 9. Yeast Hil family proteins have on average higher Ser/Thr frequencies than the rest of the proteome. Proteome-wide distribution of Thr/Ser frequencies per protein from three species, compared with the yeast Hil family proteins (*M. bicuspidata* homologs were excluded because a large number of them were incomplete). The boxes represent the interquartile range (IQR), the middle thick line the median, the whiskers the 1.5 x IQR and the dots outliers outside that range.



Supplementary Figure 10. Domain schematic for the Yeast Hil family showing rapidly evolving tandem repeat sequences in the central domain of the proteins. Same as Fig. 6A except that tandem repeats belonging to different sequence clusters as determined by XSTREAM (Newman and Cooper 2007) are shown in different colors.