Parallel Expansion and Divergence of the Hyr/Iff-like (Hil) Adhesin Family in Pathogenic Yeasts Including *Candida auris*

Rachel 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 evolved multiple times in the Saccharomycetes class. A recent example is *Candida auris*, a multidrug resistant pathogen associated with a high mortality rate and multiple hospital outbreaks. Genomic changes shared between independently evolved pathogens could reveal key factors that enable them to infect the host. One such change may be the expansion of cell wall adhesins, which mediate biofilm formation and adherence and are established virulence factors in *Candida* spp. Here we show that homologs of a known adhesin family in *C. albicans*, the Hyr/Iff-like (Hil) family, repeatedly expanded in divergent pathogenic *Candida* lineages including in *C. auris*. Evolutionary analyses reveal varying levels of selective constraint and a potential role of positive selection acting on the ligand-binding domain during the family expansion in *C. auris*. The repeat-rich central domain evolved rapidly after gene duplication, leading to large variation in protein length and β-aggregation potential, both known to directly affect adhesive functions. Within *C. auris*, isolates from the less virulent Clade II lost five of the eight Hil homologs, while other clades show abundant tandem repeat copy number variation. We hypothesize that expansion and diversification of adhesin gene families are a key step towards the evolution of fungal pathogens and that variation in the adhesin repertoire could contribute to within and between species differences in the adhesive and virulence properties.

Introduction

Candida auris is a newly emerged multidrug-resistant yeast pathogen. It 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 (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 evolutionary origin of *C. auris* as a pathogen is part of a bigger evolutionary puzzle: *C. auris* belongs to a polyphyletic group known by the genus name of *Candida*, which contains most of the human yeast pathogens. Phylogenetically, however, species like *C. albicans*, *C. auris* and *C. glabrata* belong to distinct clades with close relatives that are not or rarely found to infect humans (Fig 1A). This strongly suggests that the ability to infect humans has evolved multiple times in yeasts (Gabaldón et al. 2016). As many of the newly emerged *Candida* pathogens are resistant or can quickly evolve resistance to antifungal drugs (Lamoth et al. 2018; Srivastava et al. 2018), it is urgent to understand how yeast pathogens arose and what make them better at surviving in the host. We reason that any shared genetic changes or biological processes affected among independently derived *Candida* pathogens could reveal key factors for host adaptation and could lead to new prevention and treatment strategies.

Gene duplications and the subsequent functional and regulatory changes are a major driver in evolution (Zhang 2003; Qian and Zhang 2014; Eberlein et al. 2017). For example, this mechanism was found to underlie the independent origin of digestive RNases in Asian and African leaf monkeys (Zhang 2006), as well as the ability of insects to feed on plants that produce toxic cardenolides (Zhen et al. 2012). In support of a key role for gene duplication and sequence divergence in the emergence of yeast pathogens, a genome comparison of six *Candida* species and related low-pathogenic potential species identified a list of pathogen-enriched gene families (Butler et al. 2009). Among the top six families, three are GPI-anchored cell wall proteins – Hyr/Iff-like, Als-like and Pga30-like – that are known or suggested to act as fungal adhesins. These heavily glycosylated cell wall proteins typically have a ligand-binding domain at the N-terminus, followed by a central domain rich in tandem repeats (Fig 1B). They play key roles in adhesion to host epithelial cells, biofilm formation and iron acquisition, and are well-established virulence factors (de Groot et al. 2013; Lipke 2018). It has been suggested that expansion of cell wall protein families, 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 pathogenic *Candida* species within the Nakaseomyces genus (Gabaldón et al. 2013).

Despite the importance of adhesins in both the evolution and virulence of *Candida* pathogens, few studies have examined their evolutionary history, sequence divergence and the role of natural selection in pathogenic yeast species (Linder and Gustafsson 2008). In particular, little is known about adhesin genes in *C. auris* and their evolutionary relationship with homologs in other *Candida* species (Kean et al. 2018; Singh et al. 2019; Muñoz et al. 2021). Our goal in this study is to characterize and examine the evolutionary history and sequence divergence of adhesin genes in *C. auris* (Fig 1C, D). To identify candidate adhesins in *C. auris*, we draw on *C. albicans*, which belongs to the same CUG-Ser1 clade. Among known adhesins in *C. albicans* (Fig 1C), *C. auris* lacks the Hwp family and has only three Als or Als-like proteins, many fewer than the eight Als proteins in *C. albicans* (Fig 2A) (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 the non-pathogenic ones (Butler et al. 2009). Furthermore, 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 in this study that the Hil family has convergently expanded in *C. auris* and *C. albicans* as well as in other pathogenic *Candida* species. Sequence features and predicted effector domain structure support the majority of the yeast Hil family, including all eight members in *C. auris*, as encoding adhesins. Evolutionary analyses reveal varying levels of selective constraint and a possible role of positive selection acting on the effector domain, while rapid divergence in the repeat-rich central domain leads to large variation in length and β-aggregation potential that could affect the adhesive properties of the yeast cells and thus generates phenotypic diversity.

Results

Parallel expansion of the Hyr/Iff-like family in multiple pathogenic *Candida* lineages

The Hyr/Iff family was first identified and characterized in *Candida albicans* (Bailey et al. 1996; Richard and Plaine 2007). A defining feature of the family is its ligand-binding domain, known as Hyphal_reg_CWP (PF11765), at the N-terminus. It is followed by a variable central domain rich in tandem repeats (Boisramé et al. 2011). In a previous study, Butler *et al* used "Hyr/Iff-like" to refer to any gene sharing sequence homology in either the ligand-binding domain or the repeat domain with the Hyr/Iff genes in *C. albicans* (Butler et al. 2009). In this study we restrict the

Hyr/Iff-like (Hil) family as referring to the group of evolutionarily related proteins containing the Hyphal_reg_CWP domain at the N-terminus, thus requiring both the presence of the ligand-binding domain and also conservation of its relative position in the protein.

We identified a total of 104 Hil family homologs from 18 species in the Saccharomycetes class (Table S1). No credible hits were identified outside of Saccharomycetes, suggesting that this family is likely specific to the yeast. Notably we didn't identify any homolog in the well-studied *S. cerevisiae* or its close relatives. Although the Pfam database does contains two *S. cerevisiae* proteins in the PF11765 domain family, we found that these two proteins are not only more divergent from those in *C. auris* than homologs in the equally distant *C. glabrata*, but also have a different domain organization, with their PF11765 domains in the middle rather than at the N-terminus of the proteins (Fig S1).

To infer the evolutionary history of the Hil family, especially the history of duplications among independently evolved *Candida* pathogens, we reconstructed a phylogenetic tree based on the PF11765 domain (Fig. 2B). We found that homologs from the Clavispora and Candida genera, which include *C. auris* and *C. albicans*, respectively, formed their own groups. This suggests that the duplications in the Hil families in the two clades occurred independently. To infer the timing of the duplication and loss events, we reconciled the PF11765 domain tree with the species tree (Materials and Methods). The result suggests a duplication at the root of the CUG-Ser1 clade, followed by repeated, parallel duplications in the Candida and Clavispora genera (Fig 2C). To highlight the uneven distribution of duplications among species, we inferred the number of gains and losses on each branch in the species tree, which shows the extensive and parallel expansion of the Hil family particularly in the *albicans* and the MDR clades (Fig 2D). In the literature the *C. auris* Hil family genes have been referred to by their most closely related Hyr/Iff genes in *C. albicans* (Kean et al. 2018; Jenull et al. 2021; Muñoz et al. 2021). To avoid the incorrect implication of one-to-one orthology between the *HIL* genes in the two species, we renamed the *C. auris* Hil family genes as Hil1-Hil8 ordered by their protein length (Table S2).

Sequence features and predicted effector domain structure support *C. auris* Hil family as adhesins

Determining the adhesin status of the Hil family is important for understanding the implications of its parallel expansions. Experimental studies supported 11 of the 12 members of the Hil family proteins in *C. albicans* as adhesins (Bailey et al. 1996; Boisramé et al. 2011; Rosiana et al. 2021). Here we provide bioinformatic evidence supporting an adhesin function for all eight Hil proteins in *C. auris*. We take advantage of the characteristic domain architecture in known yeast adhesins, which consist of an N-terminal signal peptide, a ligand-binding (effector) domain, a

Ser/Thr-rich central domain with tandem repeats and β-aggregation prone sequences, and a Glycosylphosphatidylinositol (GPI) anchor at the C-terminus (Fig 3A) (de Groot et al. 2013; Lipke 2018). All eight *C. auris* Hil proteins share this domain architecture (Fig 3B) and have elevated Ser/Thr frequencies compared with the genome-wide distribution (Fig S2,3). All eight members were also predicted to be fungal adhesins by FungalRV, a support vector machine based classifier using amino acid composition and hydrophobic properties as input and showing high sensitivity and specificity in eight pathogenic fungi (Chaudhuri et al. 2011).

The structure of the effector domain in several yeast adhesin families, such as the Als, Epa and Flo families, have been solved and reveal a carbohydrate or peptide binding activity (Willaert 2018). Since an experimentally determined structure is not available for the PF11765 effector domain, we used the recently released AlphaFold2 (Jumper et al. 2021) to predict the structures of the PF11765 domains in *C. auris* Hil1 and Hil7. We chose these two because the PF11765 domain in Hil1 is representative of 6 of the 8 Hil proteins while Hil7's is the least similar in sequence to the rest (Fig S4). Both predicted structures are of high confidence and adopt a highly similar β-solenoid fold, i.e., a superhelical arrangement of repeating β-strands around a central axis, stacked into an elongated cylinder (Fig 3C, D). The β-strand-rich nature is consistent with the structurally characterized yeast adhesin effector domains, although most of them have a different, β-sandwich fold (Willaert 2018). To understand the potential function of the PF11765 domain, we searched for similar structures with known functions using the threading-based prediction server, I-TASSER (Zhang 2008). I-TASSER identified templates with good structural alignment (normalized z-scores between 1 and 2) but low sequence identity (< 20%). Remarkably, five of the six unique PDB structures in the top 10 list are from the binding domains of bacterial adhesins, such as the Serine-Rich Repeat Proteins (SRRPs) from *L. reuteri* (Fig 3E, Table 1 & S3) (Sequeira et al. 2018). Originally no yeast hits were found. This changed when a new study reported the same β-solenoid fold for two Adhesin-like wall proteins (Awp)'s effector domain from *C. glabrata* (PDB: 7O9Q, 7O9O/7O9P), which do not encode the PF11765 domain (Reithofer et al. 2021). Together, these results strongly support the ligand-binding activities for the PF11765 domain and the Hil proteins in *C. auris* as adhesins. The low sequence identity between the PF11765 domain, the bacterial adhesin binding regions and the *C. glabrata* Awp's effector domain further suggests that bacterial and yeast adhesins have convergently evolved towards a similar structure to achieve adhesion functions.

Diverged central domain may affect the adhesion function of the Hil proteins in *C. auris*

While the overall domain architecture is well conserved, the eight Hil family paralogs in *C. auris* differ significantly in length and sequence in their central domains. While the latter is not involved in ligand binding, they nonetheless play critical roles 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). Moreover, they typically encode tandem repeats and β-aggregation sequences, which 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). Hence divergence in the central domain properties has the potential to generate functional 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 to examine their similarity. We found *C. auris* Hil1 to Hil4 share a ~44 aa repeat unit, whose copy number varies from 15 to 46, which drives their difference in length (Fig 4A). Hil7 and Hil8 encode the same repeat unit but has only one copy (Fig 4B, C). By contrast, Hil5 and Hil6 encode very different, low complexity repeats with a period of 5-9 aa and between 14 to 49 copies (Fig 4D, E). These variation also affected the Ser/Thr frequencies (Fig S2).

In addition to protein length and Ser/Thr frequencies, the tandem repeat evolution also leads to differences in the β-aggregation potential by altering the number and quality of β-aggregation prone sequences. Most 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). In *C. auris* Hil1 through Hil4, however, the shared ~44 aa tandem repeat unit contains a heptapeptide ("GVVIVTT" and its variants) that is predicted to have >90% β-aggregation potential. As a result, the central domains of these proteins contain 21 to 50 highly β-aggregation-prone sequences (e.g., Hil1 shown in Fig S5). We hypothesize that the unusually high number of β-aggregation sequences in Hil1-4 and the large variation among the *C. auris* Hil proteins – only 2-4 were identified in Hil5-Hil8 – lead to diverse adhesion functions within the *C. auris* Hil family.

Intraspecific variation in Hil family size and tandem repeat copy number in *C. auris* could drive phenotypic diversity in adhesion and virulence

C. auris isolates from geographically and genetically divergent clades contain varying numbers of Hil family homologs (Muñoz et al. 2021). In particular, strains from the East Asian Clade, or Clade II, have only three of the eight members, while most strains from the other clades have eight (Muñoz et al. 2021). Our phylogenetic analysis shows that clade II strains lost Hil1-Hil4 and Hil6 (Fig S6). 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). Together, these suggest that clade II strains may have reduced adhesive capability. Interestingly, this lack of putative adhesins in Clade II coincide with the observation that >93% of Clade II isolates described in a study were associated with ear infections in contrast to invasive infections and hospital outbreaks typically caused by the other clades, and they also appear to be less resistant to antifungals (Kwon et al. 2019; Welsh et al. 2019).

Tandem repeats are prone to recombination-mediated expansions and contractions, which in turn can contribute to diversity in cell adhesive properties, as shown in *S. cerevisiae* (Verstrepen et al. 2005). Sampling nine strains in *C. auris*, we observed clade-specific variation in tandem repeat copy number in Hil1-Hil4 (Table 2). Except for one 16 aa deletion affecting one strain, all seven remaining indels correspond to one or multiples of a full repeat, consistent with their being driven by recombination between repeats (Fig S7).

Natural selection on the effector domain and the tandem repeats in *C. auris* Hil genes

Gene duplication 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). If positive selection is involved, it can lead to an elevated ratio of nonsynonymous to synonymous substitution rates dN/dS > 1 (Yang 1998). Here we ask if the ligand binding (PF11765) domain in *C. auris* Hil1-Hil8 showed any signature of positive selection during the Hil family expansion.

We first tested the hypothesis that the PF11765 domain has evolved under a constant selection strength during the expansion of the Hil family in *C. auris*. A likelihood ratio test (LRT) comparing the one-ratio model (constant selection) with the free-ratio model (varying selection at each branch) is highly significant ($2\Delta l = 446.68$, $P < 10^{-10}$ for X^2 with d.f. = 13). This suggests that selection strengths vary among lineages. The free-ratio model identified two branches with a dN/dS ratio far greater than one (ω1, 2 in Fig 5A). We tested if one or both have significantly higher dN/dS than the other branches (tests a, b and c in Table 3). The LRT results supported all three hypotheses, either tested together (a) or separately (b and c). We further asked if their dN/dS ratios are significantly greater than 1 (tests d, e and f in Table 3). Only the test with the two branches combined is significant at a 0.05 level. Two more branches showed elevated dN/dS ratios that are close to or just above 1 under the free-ratio model (labeled ω3 in Fig 5). LRT supports them being significantly different from the background dN/dS (test g, Table 3). Our results thus identified four branches with significantly elevated dN/dS over the background, with two of them showing modest evidence for dN/dS > 1, consistent with positive selection acting on the PF11765 domain. Overall, we conclude that expansion of the Hil family in *C. auris* was accompanied by relaxation of selective constraints on the PF11765 domain and may have involved episodes of positive selection driving functional divergence.

We showed previously that the central domain, especially the tandem repeats therein, evolved rapidly within the *C. auris* Hil family. Given their potential to affect the adhesin functions, we ask what types of selective forces govern the evolution of the tandem repeats. Hil1 and Hil2 duplicated recently in *C. auris* (Fig S6) and their repeats have a conserved 44 aa period (Table 2), allowing us to answer this question. Following a pioneer study by (Persi et al. 2016) on tandem repeat evolution, we estimated the pairwise dN/dS ratios between individual repeats within Hil1/Hil2 (termed "horizontal evolution") and compared them to the estimates between the repeats across the two proteins ("vertical evolution", Fig 5B). Phylogenetic tree for the repeats suggests that most of the repeats in Hil1 and Hil2 either originated after gene duplication or were subject to homogenization by gene conversion (Fig 5C). As a result, orthology between the repeats across genes is limited and difficult to determine. Thus, we inferred the selective strength for vertical evolution using pairwise dN/dS estimates between a set of 17 repeats from each of Hil1 and Hil2 (cyan lines, Fig 5B). As an alternative approach, we assumed a relatively well-aligned part of the tandem repeat region is orthologous and estimated dN/dS based on that (yellow region, Fig 5B). Both approaches yielded similar results: the distributions of dN/dS ratios within Hil1 or Hil2 are similar to each other (Fig 5D, Wilcoxon Rank Sum Test $P = 0.10$), and are significantly different (lower) than that for the inter-Hil1-Hil2 repeats (Wilcoxon Rank Sum Test $P < 0.01$). This suggest that after gene duplication, the repeats in one or both copies were under relaxed constraint or possibly positive selection, which allowed them to diverge between the two genes. Afterwards, there was increased constraint in each gene to maintain the repeats within a gene. The dN/dS ratios of the repeats either within or between the two genes are higher than those obtained for the PF11765 domain between Hil1, Hil2 and closely related MDR homologs (Fig 5D), suggesting that the repeats in general evolved under weaker selective constraint than did the PF11765 domain.

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

Above we focused on the Hil family in *C. auris* and provided a detailed picture of the adhesin features and sequence divergence after duplication. Here we apply these analyses to the entire Hil family in yeasts. We found that 92/104 homologs were predicted to be fungal adhesins by FungalRV, and 97 and 89 were predicted to have a signal peptide and GPI-anchor, respectively (Fig S8A), consistent with most of the yeast adhesins being GPI-anchored cell wall proteins (Lipke 2018). 76 of the 104 Hil homologs passed all three tests. Moreover, all but five homologs encode tandem repeats in their central domain, with proteins longer than 1500 aa having a significantly higher proportion of their central domain consisting of tandem repeats (Fig S8B). Hil homologs also have a higher serine and threonine content compared with the proteome-wide distribution (Fig S8C). All of them have at least one β-aggregation prone sequence. Finally, structural predictions for the PF11765 domain in three Hil proteins from *C. albicans*, *C. glabrata* and *K. lactis* all showed a similar β-solenoid fold as predicted for *C. auris* Hil1 and Hil7 and shared with the bacterial SRRP adhesins (Fig S9). Together, these lines of evidence suggest that the majority of the yeast Hil family encode fungal adhesins.

Similar to our findings in *C. auris*, the yeast Hil family as a whole exhibits large variation in protein length and sequence properties within their central domain (Fig 6). For protein length, the non-PF11765 portion of these proteins have a mean and standard deviation of 936.8±725.1 aa and a median of 650.5 aa (Fig 6A). This variation in protein length is almost entirely driven by the tandem repeats (Fig 6B, linear regression slope = 0.996, r^2 = 0.76). Not only do the tandem repeats vary in copy number, but the underlying sequences also diverged rapidly (Fig S10, Table S4). This leads to large variation in sequence properties such as β-aggregation potential (Fig 6C). A subset of Hil homologs consisting of *C. auris* Hil1-4 and their closely related proteins in the MDR clade are unique even within the family: they are longer than the other Hil homologs (1592 vs. 918.5 aa in median length) and also have more TANGO positive motifs (22 vs 4 in median number of total hits). A curious and distinct feature of the TANGO motifs in this group is that they are regularly spaced as a result of the motif being part of the repeat (median absolute deviation, or MAD, of distances between adjacent strong TANGO "hits" less than 5 aa, Fig. 6D). The heptapeptide "GVVIVTT" and its variants account for 61% of all hits in this subset and are not found in the other Hil homologs (Table S5).

The yeast Hil family genes are preferentially located near chromosome ends

Several well-characterized yeast adhesin families, such as the Epa family in *C. glabrata* and the Flo family in *S. cerevisiae*, are enriched in the subtelomeres (Teunissen and Steensma 1995; De Las Peñas et al. 2003). This region is associated with high rates of SNPs, indels and copy number variations, and can undergo ectopic recombination that can lead to the spread of genes between chromosome ends or their losses (Mefford and Trask 2002; Anderson et al. 2015). We found that the yeast Hil family genes are frequently located near the chromosome ends as well (Fig S11). To test if this trend is significant, we compared their chromosomal locations with the background gene density distribution in six species whose genomes are assembled to a chromosomal level (Table S6, Materials and Methods). We found the Hil family genes are indeed enriched at the chromosome ends (Fig. 7A, B). A goodness-of-fit test confirmed that the difference between the distribution of chromosomal locations of the Hil family and the genome background is significant ($P = 3.6 \times 10^{-6}$). It has been shown that ectopic recombination between subtelomeres can lead to the spread and amplification of gene families (Anderson et al. 2015).

We thus 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 (Fig 7C).

Discussion

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. In this study, we resolved a detailed evolutionary history for the Hyr/Iff-like (Hil) family and characterized its sequence divergence and the selection forces involved. Our results support the previous finding that adhesin families are enriched in pathogenic yeasts (Fig 2A). Phylogenetic analysis convincingly showed that this correlation resulted from convergent expansions, with most of the duplications occurring in the albicans clade and the Multi-Drug Resistant (MDR) clade in two separate genera (Fig 2D).

The Hil family was experimentally studied in *C. albicans* (Bailey et al. 1996; Luo et al. 2010; Boisramé et al. 2011), revealing 11 of its 12 members as GPI-anchored cell wall proteins with a potential role in adhesion. Similar evidence is lacking for family members in other yeasts. We showed that ~75% of all Hil proteins, including all eight members in *C. auris*, are predicted to be GPI-anchored cell wall proteins and pass a fungal adhesin predictor's (FungalRV) cutoff, supporting the adhesin status for the Hil family in general. We also used AlphaFold2 to make high-confidence predictions for the effector domain structure in several distantly related Hil proteins, all of which showed the same β-solenoid fold (Fig 3C-E, S8). This structure is highly similar to the binding region of some bacterial adhesins, e.g., the Serine Rich Repeat Protein (SRRP) in *L. reuteri* (Sequeira et al. 2018) as well as two newly reported yeast adhesin effector domains (Reithofer et al. 2021). The cross-kingdom similarity in the adhesin effector domain structure is intriguing in several ways. First, it suggests convergent evolution in bacteria and yeasts. Second, what's known about the structure-function relationship in bacteria can provide insight into the PF11765 domain in yeast. Notably, *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).

Sequence divergence after gene duplication allows for sub- or neo-functionalization that fuels evolution (Zhang 2003; Innan and Kondrashov 2010; Eberlein et al. 2017). Using *C. auris* as a focal species, we found that while the PF11765 domain in its *HIL* genes evolved under purifying selection in general (dN/dS < 0.2), four branches showed significantly higher dN/dS ratios, including two with modest evidence for a dN/dS > 1, suggesting positive selection in addition to relaxed selective constraints (Fig 5A, Table 3). The implication is that changes in the effector domain sequence could affect the specificity or affinity for its substrates, which in turn could impact the adhesive properties of the cell. Experiments to characterize the binding affinity and substrate specificity of the eight Hil proteins in *C. auris* will be highly desired. Compared to the conserved effector domain, the central domain of the Hil family evolved much more rapidly after gene duplication, generating large variation in protein length and β-aggregation potential (Fig 3, 6). Evolutionary analyses comparing the repeat sequences in the recently duplicated Hil1 and Hil2 showed that 1) the tandem repeats were also subject to purifying selection, albeit to a less extent than the PF11765 domain; 2) most of the repeats in the two genes likely originated after gene duplication, underscoring their dynamic nature; 3) the dN/dS ratios are slightly higher for repeats across the two genes than within each gene, consistent with a period of relaxed constraint after gene duplication. Although a role for positive selection cannot be ruled out. Together, our analyses painted a detailed evolutionary picture for how repeats originate, evolve and are selectively maintained.

Variations in protein length and β-aggregation potential resulting from the central domain divergence could directly impact the adhesion functions (Verstrepen et al. 2005; Alsteens et al. 2010; Ramsook et al. 2010; Boisramé et al. 2011; Lipke et al. 2012). In this regard, we found *C. auris* Hil1-4 and the closely related MDR homologs to be unusual as they have as many as 50 β-aggregation prone sequences in contrast to 1-3 in known yeast adhesins (Ramsook et al. 2010). This raises the question of whether they possess special adhesive properties. In addition to sequence divergence between homologs, we also identified intraspecific variation in the size and tandem repeat copy number of the Hil family. It has been shown previously that the Clade II strains in *C. auris* lack five of the eight Hil genes (Muñoz et al. 2021). We showed that this is due to gene loss (Fig S6). Interestingly, Clade II strains are unique among *C. auris* strains in that they are mostly associated with ear infections rather than hospital outbreaks as the other clades do (Kwon et al. 2019; Welsh et al. 2019). Since they also lack a *C. auris* specific GPI-anchored cell wall protein family (Muñoz et al. 2021), we hypothesize that Clade II strains have weaker adhesive abilities, which may be a cause or consequence of their distinct niche preference. We also found tandem repeat copy number variations in Hil1-Hil4 among clade I, III and IV strains in *C. auris*. As shown experimentally for the *S. cerevisiae* Flo family, adhesin protein length is strongly correlated with the adhesive properties and the flocculation and biofilm formation capabilities (Verstrepen et al. 2005). Thus, Hil protein length variations in *C. auris* could further contribute to diversity in its adhesive properties and virulence.

Finally, we found that the Hil family genes are preferentially located near chromosomal ends in the species examined (Fig 7), similar to previous findings for the Flo and Epa families (Teunissen and Steensma 1995; De Las Peñas et al. 2003). This location bias can be both a cause and consequence of the family expansion, as it is known that subtelomeres are subject to ectopic recombination that can lead to the spread of gene families between chromosome ends (Mefford and Trask 2002; Anderson et al. 2015). In addition to a higher rate of gene gains and losses, there are two other consequences for the Hil family being located in the subtelomeres: 1) the higher rates of mutations and structural variations associated with the subtelomeres could drive rapid diversification of the adhesin gene family (Snoek et al. 2014; Xu et al. 2021); 2) gene expression in the subtelomere is subject to epigenetic silencing, 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* and leads to hyperadherent phenotypes in *C. glabrata* (Halme et al. 2004; Castaño et al. 2005).

Together, our results provide a detailed phylogenetic analysis for a putative adhesin family in the Saccharomycetes, supporting the hypothesis that parallel expansions and the ensuing diversification of adhesins are a key step towards the evolution of yeast pathogens. Our results point to possible functional divergences between and within species in terms of adhesive properties, particularly in the emerging, multi-drug resistant species *C. auris*, which could have significant impact on their virulence profiles.

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
FungalRV	(Chaudhuri et al. 2011)	http://fungalrv.igib.res.in/
SignalP 5.0	(Almagro Armenteros et al. 2019)	http://www.cbs.dtu.dk/services/SignalP/
PredGPI	(Pierleoni et al. 2008)	http://gpcr.biocomp.unibo.it/predgpi/
hmmscan	(Potter et al. 2018)	https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan
XSTREAM	(Newman and Cooper 2007)	https://amnewmanlab.stanford.edu/xstream/download.jsp
EMBOSS v6.6.0.0	(Rice et al. 2000)	http://emboss.open-bio.org/
TANGO v2.3.1	(Fernandez-Escamilla et al. 2004)	http://tango.crg.es/
JDotter	(Brodie et al. 2004)	https://4virology.net/virology-ca-tools/jdotter/
Clustal Omega v1.2.4	(Sievers et al. 2011)	http://www.clustal.org/omega/
Jalview v2.11.1.4	(Waterhouse et al. 2009)	https://www.jalview.org/
BLAST+ v2.12.0	(Camacho et al. 2009)	https://blast.ncbi.nlm.nih.gov/
AlphaFold2	(Jumper et al. 2021)	https://github.com/sokrypton/ColabFold (links to DeepMind Google Colab Notebook)
SwissModel	(Waterhouse et al. 2018)	https://swissmodel.expasy.org/
I-TASSER	(Zhang 2008)	https://zhanggroup.org/I-TASSER/
PyMol v2.5.2	(Schrödinger, LLC 2021)	https://pymol.org/
RAxML v8.2.12	(Stamatakis 2014)	https://cme.h-its.org/exelixis/web/software/raxml/
GeneRax v2.0.1	(Morel et al. 2020)	https://github.com/BenoitMorel/GeneRax
FigTree v1.4.4	NA	http://tree.bio.ed.ac.uk/softwar

		e/figtree/
Notung v2.9	(Chen et al. 2000)	http://www.cs.cmu.edu/~durand/Notung/
PAML v4.9e	(Yang 2007)	http://abacus.gene.ucl.ac.uk/software/paml.html
bedtools v2.30.0	(Quinlan and Hall 2010)	https://bedtools.readthedocs.io/en/latest/index.html
PAL2NAL.pl	(Suyama et al. 2006)	http://www.bork.embl.de/pal2nal/
R v4.1.0	(R Core Team)	https://cran.r-project.org/
R package - ggtree v3.2.1	(Yu 2020)	https://github.com/YuLab-SMU/ggtree
R package - treeio v1.18.1	(Wang et al. 2020)	https://github.com/YuLab-SMU/treeio
R package - rentrez v1.2.3	(Winter 2017)	https://github.com/ropensci/rentrez
RStudio v1.4	(RStudio Team 2021)	https://www.rstudio.com/
Custom R, Python and shell scripts	This study	https://github.com/binhe-lab/C037-Cand-auris-adhesin

METHOD DETAILS

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

To identify the Hyr/Iff-like (Hil) proteins in *C. auris*, we used the Hyphal_reg_CWP domain from Hil1 of B11221 as the query and searched against the annotated protein sequences from the representative strains in Clade I to Clade IV (B8441, B11220, B11221, B11243) using blastp (v2.12.0, "-max_hsps = 1"). To identify the Hil family proteins in yeasts and beyond, we used the same query as above and searched the RefSeq protein database with an E-value cutoff of 1x10^{-5}, a minimum query coverage of 50% and with the low complexity filter on. All 189 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. We refined the remaining list of sequences by removing the following species, which were already represented by well-studied relatives in the list: *Metschnikowia bicuspidata var. Bicuspidata*, *Debaryomyces fabryi*, *Suhomyces tanzawaensis*, *Candida orthopsilosis*, *Meyerozyma guilliermondii*, *Yamadazyma tenuis*, *Diutina rugosa*, *Kazachstania africana*, *Kazachstania naganishii*, *Naumovozyma dairenensis* and *Cyberlindnera jadinii*. We further excluded those that were 500 aa or shorter (notably the fission yeast hit is 339 aa). This was based on studies of the Epa family in *C. glabrata* and the Hyr/Iff family in *C. albicans* showing that a critical length is required for the adhesin function (Frieman et al. 2002; Boisramé et al. 2011). The 27 sequences that were removed by the length criterion were primarily from two species: *C. parapsilosis* (10) and *S. stipitis* (12) (Table S7). In total 95 sequences were left after both filtering steps.

The RefSeq database lacks many yeast species such as those in the Nakaseomyces genus, which includes multiple *Candida* pathogens. We thus searched two additional yeast-specific databases: FungiDB (Basenko et al. 2018) and Genome Resources for Yeast Chromosomes (GRYC, http://gryc.inra.fr/). Using the same criteria, we recovered five and four additional sequences, resulting in a final dataset of 104 homologs from 18 species.

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

To infer the evolutionary history of the Hil family, which is characterized by its single effector domain, the PF11765 domain, we reconstructed a phylogenetic tree based on the alignment of that domain. First, the N-terminal 500 amino acid sequences for each Hil family protein were extracted, which included the PF11765 domain. These sequences were then aligned using Clustal Omega with the parameter {--iter=5}. The alignment was manually inspected and the first 480 columns were determined to contain the PF11765 domain and thus used for gene tree reconstructions. RAxML v8.2.12 was compiled and run on the University of Iowa ARGON server with the following parameters on the alignment: “mpirun raxmlHPC-MPI-AVX -f a -x 12345 -p 12345 -# 500 -m PROTGAMMAAUTO”. The resulting tree was manually inspected in FigTree (v1.4.4). To infer the history of duplications and losses, the gene tree was reconciled with a species tree based on the literature (Muñoz et al. 2018; Shen et al. 2018) using Notung v2.9 (Chen et al. 2000). To do so, the protein names in the gene tree were edited to include the species name as a postfix. In Notung, we first ran a rooting analysis which, in agreement with our expectation, identified the branch that separated the Saccharomycetaceae sequences from the CUG-Ser1 sequences as the best root choice. The reconciled tree was then rearranged with an edge weight threshold of 80.0, which allowed branches with less than 80% rapid bootstrapping support to be swapped. All rearrangements were ranked by the total event score, which is a weighted sum of penalties for duplications (1.5) and losses (1.0). The rearrangement with the lowest total event score was chosen as the most likely tree. As the branch length values for the swapped branches were no longer meaningful, the final tree was represented as a cladogram. Tree annotation and visualization were done in R using the treeio and ggtree packages (Wang et al. 2020; Yu 2020).

To refine the phylogenetic tree for the Hil family in *C. auris* and infer gains and losses within the species, we identified orthologs of the Hil genes in representative strains of 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 an outgroup *D. hansenii* were also included. Gene tree was constructed as described above. To root the tree, we first inferred a gene tree without including 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 --max-spr-radius 5} (Morel et al. 2020). The inferred gene tree was used as the starting tree and a "species" tree that depicts the relationship between the strains of *C. auris* and the three other species was based on (Muñoz et al. 2018).

Prediction of adhesin-related sequence features

1) Signal Peptide was predicted using the SignalP 5.0 server, with the "organism group" set to Eukarya (Almagro Armenteros et al. 2019). The server reported the proteins that had predicted signal peptides. No further filtering was done. **2)** GPI-anchor was predicted using PredGPI (Pierleoni et al. 2008) using the General Model. The server reports the false positive rate and predicted omega-site for each input protein. We defined proteins with a false positive rate of 0.01 or less as containing a GPI-anchor. **3)** Pfam domains in each of the proteins, including the Hyphal_reg_CWP domain, were identified using the hmmscan (Potter et al. 2018). **4)** Tandem repeats were identified using XSTREAM (Newman and Cooper 2007) 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)** Serine and Threonine content in proteins were quantified using freak from the EMBOSS suite, using a sliding window of 100 aa, with a step size of 10 aa (Rice et al. 2000). **6)** 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"} (Fernandez-Escamilla et al. 2004). **7)** Lastly, FungalRV, a Support Vector Machine based fungal adhesin predictor, was used to evaluate all Hil family proteins (Chaudhuri et al. 2011). Proteins passing the software recommended cutoff of 0.511 were considered positive.

Species 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 and a custom Python script was used to count the frequency of serine and threonine residues. The assembly information for the species is in Table S6 and the script is available in the project GitHub repository.

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. Threading-based prediction and identification of structures with similar folds were performed with the I-TASSER server (Zhang 2008). Model visualization and annotation were done in PyMol v2.5.2 (Schrödinger, LLC 2021). Secondary structure prediction for *C. auris* Hil1's central domain was performed using PSIPred (Buchan and Jones 2019).

Dotplot, identification and annotation of sequence variations among *C. auris* Hil genes

To determine the self-similarity and similarity between the eight *C. auris* Hil proteins, we made dot plots using JDotter (Brodie et al. 2004). The window size and contrast settings were labeled in the legends for the respective plots. To visualize the length polymorphism among *C. auris* Hil1 alleles, the multiple sequence alignment was created using Clustal Omega (Sievers et al. 2011) and annotated using Jalview 2 (Waterhouse et al. 2009).

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. The accession numbers can be found in (Muñoz et al. 2021). We used the amino acid sequences for Hil1-Hil4 from the strain B8441 as query and searched against the nucleotide sequences using tblastn with the following parameters {-db_gencode 12 -evalue 1e-150 -max_hsps 2}. Orthologs in each strain were manually curated based on the blast hits to either the PF11765 domain alone or the entire protein query. All Clade II strains are missing Hil1-Hil4. Several strains in Clade I, III and IV were found to lack one or more Hil proteins (Table 2). But upon further inspection, it was found that they have significant tblastn hits for part of the query, e.g., the central domain, and the hits are 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 or not in those strains.

Estimation of dN/dS ratios and testing branch and site models of Hil gene evolution

To test whether there has been relaxed selective constraint or even positive selection acting on the PF11765 domain during the expansion of the Hil family in *C. auris*, we used the "codeml" program in PAML (v4.9e) (Yang 2007) to fit and compare a series of "branch models" (Table S8). The following parameters were used: {seqtype = 1, CodonFreq = 1, model = variable, NSsites = 0, code = 8, fix_kappa = 0, kappa = 2, fix_omega = 0/1, omega = 0.4/1, cleandata = 0}, among which "model", "fix_omega" and "omega" vary among the different models. In the main text, we presented results obtained with "CodonFreq = 1" (F1x4), where the equilibrium codon frequencies were estimated based on the average nucleotide frequencies regardless of the codon position. To determine if the results were robust to how codon frequencies were estimated, we repeated the analysis with "CodonFreq = 0" (Fequal, assuming equal frequency for all 61 codons) and "CodonFreq = 2" (F3x4, codon frequencies estimated from the nucleotide frequencies at the three codon positions). The result with "CodonFreq = 0" is nearly identical to those with the results in the main text. However, the result obtained with "CodonFreq = 2" identified different branches as having elevated dN/dS ratios (Fig S12). Under this model, the dS estimates for some branches were >30 substitutions per synonymous site, with a total tree length - defined as the number of nucleotide substitutions per codon - being 100, compared with 15 and 10 under the F1x4 and the Fequal model, respectively. These unusually large estimates led us to question the validity of the F3x4 model fits to our dataset. We noticed that in our data the third codon position is rich in C/T (72%, vs 37% and 55% at the first and second positions) and has very few A's (<10%), which may be the cause for the unusual dS estimates.

To estimate the pairwise dN/dS ratios between repeats either within or across Hil1 and Hil2 in *C. auris*, we used the "yn00" program in PAML (v4.9e), which implements the method described in (Yang and Nielsen 2000). The following parameters were used: {icode = 8, weighting = 1, common3x4 = 1}. The repeats themselves in the two genes were identified using XSTREAM as described above and their sequences were manually extracted with the help of the "getfasta" tool in the BEDtools suite (Quinlan and Hall 2010). In both this and the above analysis, the coding sequence alignment files were prepared using PAL2NAL.pl (Suyama et al. 2006) with the protein sequence alignment and nucleotide sequence files as input. To test for differences in the mean of the distribution between the intra- and inter-gene pairwise dN/dS estimates, we used two-tailed Wilcoxon Rank Sum tests.

Chromosomal locations of Hil family genes

Of the 18 species, seven had been assembled to a chromosomal level and are suitable for determining the chromosomal locations of the Hil family genes (Table S6), i.e., *C. albicans*, *C. dubliniensis, C. glabrata, D. hansenii, K. lactis, N. castellii* and *S. stipitis*. *C. dubliniensis* was removed because it is closely related to *C. albicans* and our phylogenetic analysis showed that most of the Hil family genes in the two species share their duplication history. Similarly, we removed *N. castellii*, which is redundant with *K. lactis*. We note that while the *C. auris* RefSeq Assembly (B11221) is still at a scaffold level, a recent study showed that seven of its longest scaffolds are chromosome-length, thus allowing the mapping of scaffolds to chromosomes (Muñoz et al. 2021, Supplementary Table 1). We thus included *C. auris* in the downstream analysis. To determine the chromosomal locations of the Hil homologs in these six species, we used Rentrez v1.2.3 (Winter 2017) in R to query the NCBI databases with their protein IDs (scripts available in the project GitHub repository). To calculate the background gene density on each chromosome, we downloaded the feature tables for the six genomes from NCBI and calculated the location of each gene as its start coordinate divided by the chromosome length. To compare the chromosomal locations of Hil family genes to the genome background, we divided each chromosome into five equal-sized bins based on the distance to the nearest chromosome end and calculated the proportion of genes residing in each bin either for the Hil family or for all protein coding genes. To determine if the two distributions differ significantly from one other, we performed a goodness-of-fit test using either a Log Likelihood Ratio (LLR) test or a Chi-Square test, as implemented in the XNomial package in R (Engels 2015). The LLR test is generally preferred and its *P*-value is reported in the results.

Acknowledgement

We thank the members of the Gene Regulatory Evolution lab for discussions. Dr. Bin Z. He is supported by NIH R35GM137831. Lindsey Snyder was supported by the NIH Predoctoral Training grant T32GM008629. Rachel 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.

Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37:420–423.

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.

Basenko EY, Pulman JA, Shanmugasundram A, Harb OS, Crouch K, Starns D, Warrenfeltz S, Aurrecoechea C, Stoeckert CJ, Kissinger JC, et al. 2018. FungiDB: An Integrated Bioinformatic Resource for Fungi and Oomycetes. *J. Fungi Basel Switz.* 4:E39.

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.

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

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.

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.

Chen K, Durand D, Farach-Colton M. 2000. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* 7:429–447.

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.

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

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.

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.

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.

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.

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

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/

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.

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

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.

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.

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.

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.

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

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.

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

Persi E, Wolf YI, Koonin EV. 2016. Positive and strongly relaxed purifying selection drive the evolution of repeats in proteins. *Nat. Commun.* 7:13570.

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.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842.

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.

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/

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.

Snoek T, Voordeckers K, Verstrepen KJ. 2014. Subtelomeric Regions Promote Evolutionary Innovation of Gene Families in Yeast. In: Louis EJ, Becker MM, editors. Subtelomeres. Berlin, Heidelberg: Springer. p. 39–70. Available from: https://doi.org/10.1007/978-3-642-41566-1_3

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.

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.

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.

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, et al. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46:W296–W303.

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.

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

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 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.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17:32–43.

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. 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nat. Genet.* 38:819–823.

Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 9:40.

Zhen Y, Aardema ML, Medina EM, Schumer M, Andolfatto P. 2012. Parallel Molecular Evolution in an Herbivore Community. *Science* 337:1634–1637.

Table 1. Top structural templates for *C. auris* Hil PF11765 domains

PDB Hit	Species	Organism	Protein / Domain	Function
5ke1A	*Shigella flexneri*	gram-negative bacterium	IcsA/VirG passenger-domain	adhesin and actin-polymerizing factor
5nxkA	*Limosilactobacillus reuteri*	gram-positive bacterium	Serine-Rich Repeat Protein binding domain	adhesin
3ogzA	*Leishmania major*	trypanosomes	UDP-sugar pyrophosphorylase	sugar pyrophosphorylase with broad substrate specificity
5ny0A	*Limosilactobacillus reuteri*	gram-positive bacterium	Serine-Rich Repeat Protein binding domain	adhesin
4kh3A	*Escherichia coli*	gram-negative bacterium	bacterial self-associating protein	self-association and cell aggregation
6n2bA	*Caldicellulosiruptor kristjanssonii*	gram-positive bacterium	Tapirin C-terminal domain	cellulose-binding, adhesion

Table 2. Intraspecific variation in tandem repeat copy number in *C. auris* Hil1-4

	Tandem repeat: copy number (repeat period)			
Clade / Strain	Hil1	Hil2	Hil3	Hil4
Clade I				
B8441	46 (44)	21 (44)	16 (51)	15 (47)
B11205	46 (44)	21 (44)	16 (51)	15 (47)
B13916	46 (44)	21 (44)	16 (51)	15 (47)
Clade III				
B11221	48 (44)	20* (44)	16 (51)	NA
B17721	48 (44)	21 (44)	16 (51)	14 (48)
B12037	48 (44)	NA	16 (51)	14 (48)
B12631	NA	NA	16 (51)	14 (48)
Clade IV				
B12342	48 (43)	15 (44)	11 (51)	15 (47)
B11245	NA	15 (44)	11 (51)	15 (47)

* deletion of 16 aa, not a full repeat
NA: homolog not identified in the strain's genome by tblastn

801 Table 3. Likelihood ratio tests for different dN/dS ratios

Alternative Hypothesis	Null Hypothesis	Assumption Made	$2\Delta\ell$	Models Compared
a. (ω1=ω2) ≠ ω0	(ω1 = ω2) = ω0	ω1 = ω2	24.74***	A and D
b. ω1 ≠ ω0	ω1 = ω0	ω2 = ω0	17.92***	A and B
c. ω2 ≠ ω0	ω2 = ω0	ω1 = ω0	7.18**	A and C
d. (ω1=ω2) > 1	(ω1 = ω2) ≤ 1	ω1 = ω2	4.2*	D and G
e. ω1 > 1	ω1 ≤ 1	ω2 = ω0	2.68	B and E
f. ω2 > 1	ω2 ≤ 1	ω1 = ω0	1.74	C and F
g. ω3 ≠ ω0	ω3 = ω0	(ω1 = ω2) free	47.14***	D and H

802 Designations for the different ω values can be found in Fig 5A; $2\Delta\ell$ = 2x log likelihood difference between the two
803 models being compared, where the models and their parameter estimates can be found in Table S8. Tests significant
804 at 0.001, 0.01 and 0.05 levels are labeled with three, two and one asterisk(s), with *P*-values based on a χ^2 distribution
805 with 1 degree of freedom.

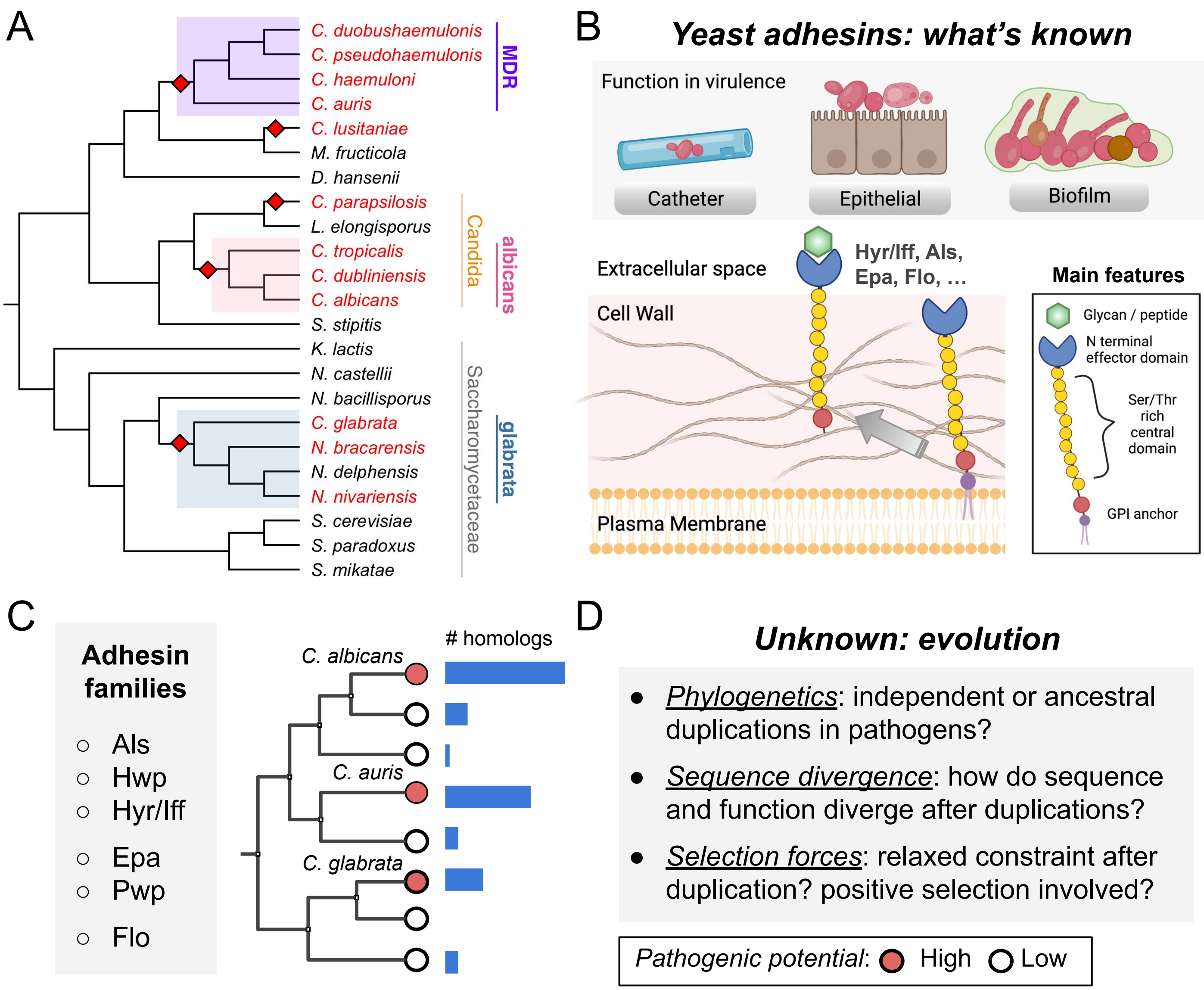
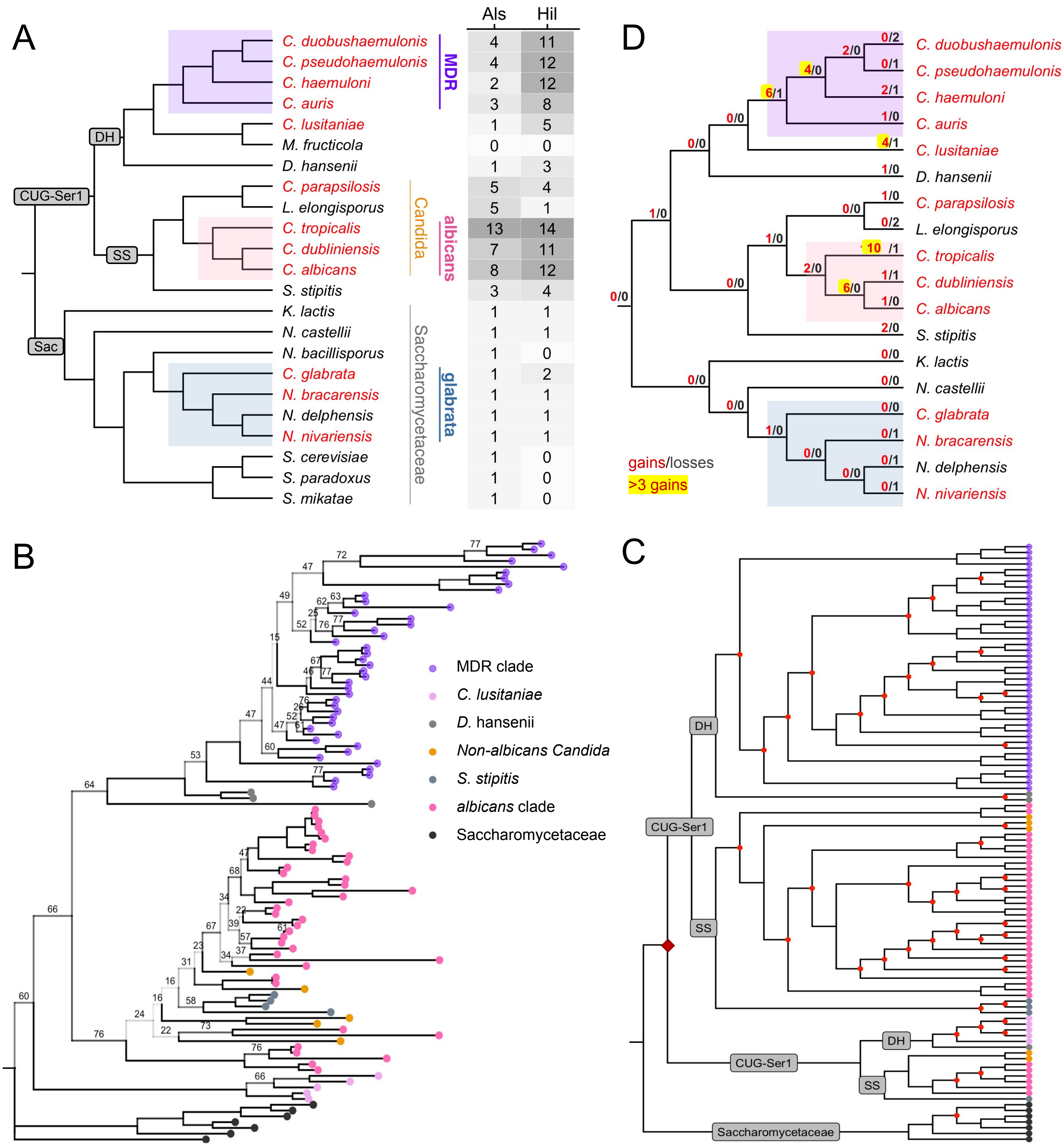


Figure 1. Multiple origins of yeast pathogens and evolution of yeast adhesin families. (A) Species phylogeny suggesting multiple origins of yeast pathogens. Species known to be pathogenic are in red and species never or rarely identified as pathogens are in black. Diamonds represent potential origination of pathogenesis, which are enriched in the highlighted *glabrata*, *albicans* and multidrug-resistant (MDR) clades. (B) As cell-wall proteins, yeast adhesins are initially inserted into the plasma membrane; most are then cleaved at the C-terminal GPI-anchor, the remnant of which allow them to be covalently linked to the β-1,6-glucan in the cell wall. The central stalk (yellow circles) is glycosylated at the Ser/Thr residues, which enables it to adopt a rigid, rod-like shape that helps to push out the N-terminal effector domain. The latter binds glycan or peptide substrates and mediates adhesion to other yeasts, host epithelium or inanimate surfaces. Drawing partly based on (Verstrepen and Klis 2006) and created with BioRender.com (C) Left: examples of known yeast adhesin families in *C. albicans* (first threes), *C. glabrata* (middle two) and *S. cerevisiae* (last). Right: a species tree showing the larger size of an adhesin family in the pathogenic species. (D) The evolutionary questions to be addressed in this study. Full species names in (A): *Candida duobushaemulonis, Candida pseudohaemulonis, Candida haemuloni, Candida auris, Clavispora lusitaniae, Metschnikowia fructicola, Debaryomyces hansenii, Candida parapsilosis, Lodderomyces elongisporus, Candida tropicalis, Candida dubliniensis, Candida albicans, Scheffersomyces stipitis, Kluyveromyces lactis, Naumovozyma castellii, Nakaseomyces bacillisporus, Candida glabrata, Nakaseomyces bracarensis, Nakaseomyces delphensis, Nakaseomyces nivariensis, Saccharomyces cerevisiae, Saccharomyces paradoxus, Saccharomyces mikatae*



Species	Als	Hil
C. duobushaemulonis	4	11
C. pseudohaemulonis	4	12
C. haemuloni	2	12
C. auris	3	8
C. lusitaniae	1	5
M. fructicola	0	0
D. hansenii	1	3
C. parapsilosis	5	4
L. elongisporus	5	1
C. tropicalis	13	14
C. dubliniensis	7	11
C. albicans	8	12
S. stipitis	3	4
K. lactis	1	1
N. castellii	1	1
N. bacillisporus	1	0
C. glabrata	1	2
N. bracarensis	1	1
N. delphensis	1	1
N. nivariensis	1	1
S. cerevisiae	1	0
S. paradoxus	1	0
S. mikatae	1	0

Figure 2. Parallel expansion of the Hil family in independently derived pathogenic *Candida* lineages. Legend on next page

Figure 2. Parallel expansion of the Hil family in independently derived pathogenic *Candida* lineages. (A) Same species tree as in Fig 1A, with gray labels in the inner nodes corresponding to those in panel C. The size of two adhesin families found in both *C. albicans* and *C. auris* are shown. (B) Maximum likelihood tree based on the binding domain of the Hil family is shown as a phylogram, rooted on the Saccharomycetaceae group. Branches with lower rapid bootstrap support by RAxML are shown as semi-transparent lines; bootstrap values lower than 80% are labeled. (C) Reconciled gene tree shown in cladogram. Gray labels highlight the important clades, including the outgroup of Saccharomycetaceae, the two CUG-Ser1 groups following an ancient duplication (red diamond) and within each branch, the Candida and Clavispora sequences labeled by their respective outgroups, *D. hansenii* (DH) and *S. stipitis* (SS). Inferred duplication events are labeled with a red circle, except for the CUG-Ser1 duplication mentioned above. (D) Species tree showing the inferred number of duplications (red) and losses (gray). Three or more duplications are highlighted in yellow. Species with zero Hil family homologs are not shown.

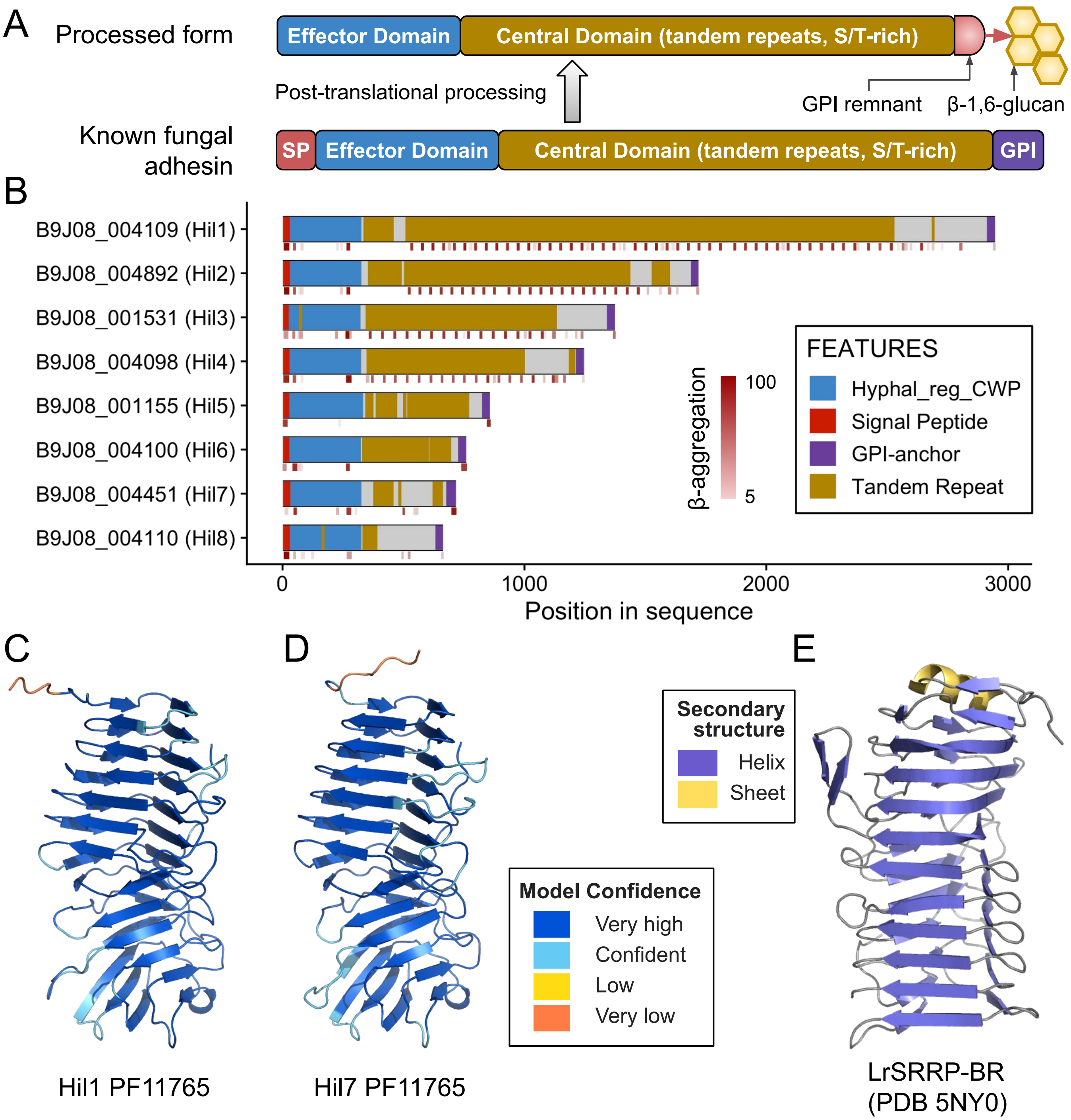


Figure 3. Domain architecture and predicted effector domain structures support *C. auris* Hil proteins as adhesins. (A) Diagram depicting a typical yeast adhesin's domain organization, before and after the post-translational processing. Adapted from (de Groot et al. 2013). (B) Domain features of the eight Hil proteins in *C. auris* (strain B8441). Gene IDs and names designated in this study 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) and (D) are AlphaFold2 predicted structures of the PF11765 domains from Hil1 and Hil7. Colors represent the local confidence score (pLDDT). (E) Experimentally determined structure of the Binding Region of the Serine-Rich-Repeat-Protein (SRRP-BR) from *L. reuteri*. Colors represent the secondary structure assignments.

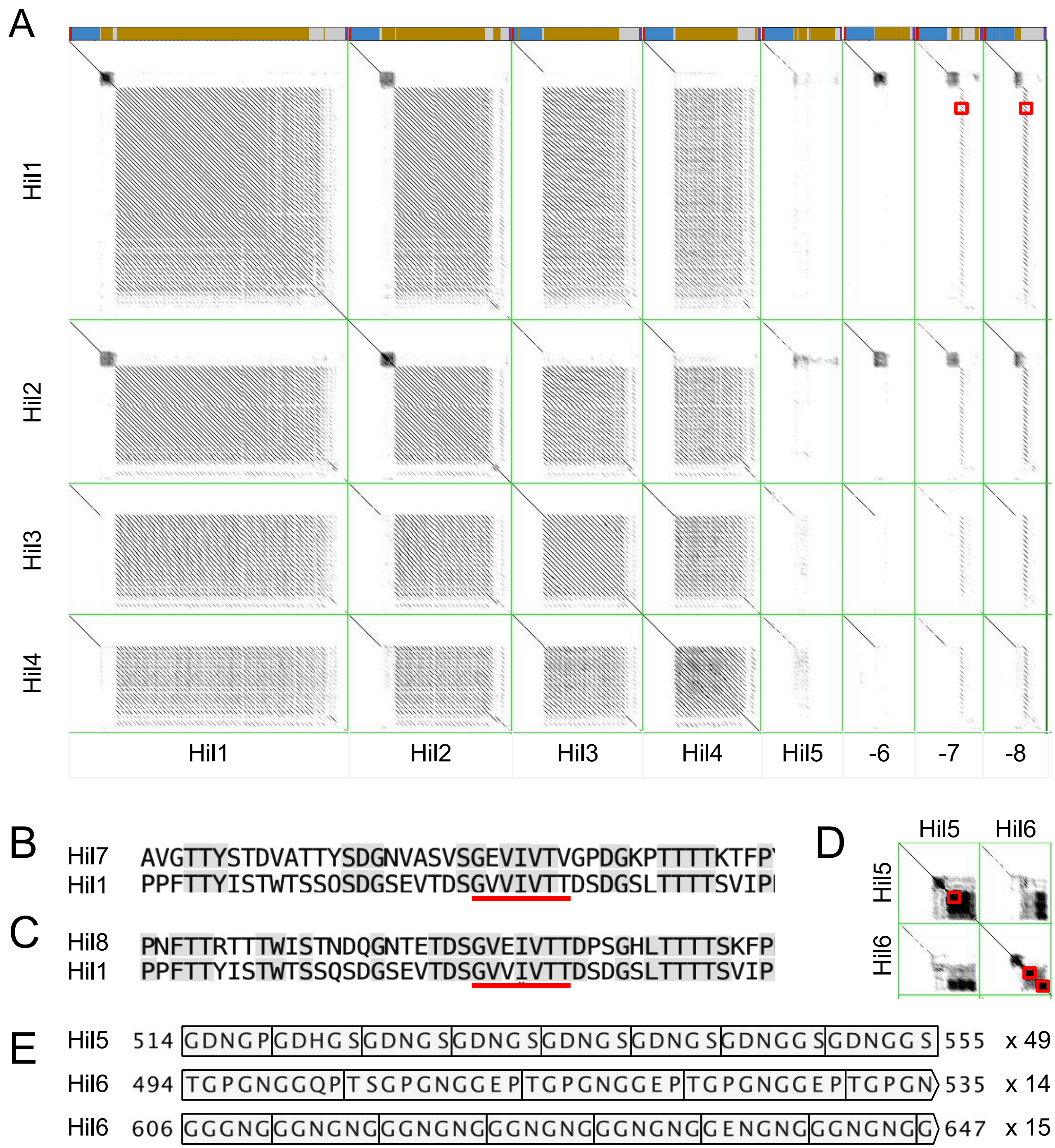


Figure 4. Dotplot shows the tandem repeat structure within and similarity between *C. auris* Hil proteins. (A) Dotplot (JDotter, Brodie et al 2004) with a sliding window of 50 aa and Grey Map set to 60-245 (min-max). Hil1-4 are compared to all eight Hil proteins including themselves. A schematic was included for each protein on the top (colors same as in Fig 3). The regions highlighted by the red boxes in row 1 are shown as sequence alignment in (B) and (C) to demonstrate the presence of a single copy of the repeat in Hil7, 8. Shadings indicate sequence similarity and the red underlines highlight the predicted β-aggregation prone sequence. (D) Dotplot between Hil5 and Hil6 with the same settings as in (A), showing the low complexity repeats unique to these two. Regions within the three red boxes are shown in (E), with limits shown on both ends of the sequences. The rectangles delineate individual repeats, with the copy numbers shown to the right. The last copy, when truncated, is indicated by a pointed shape.

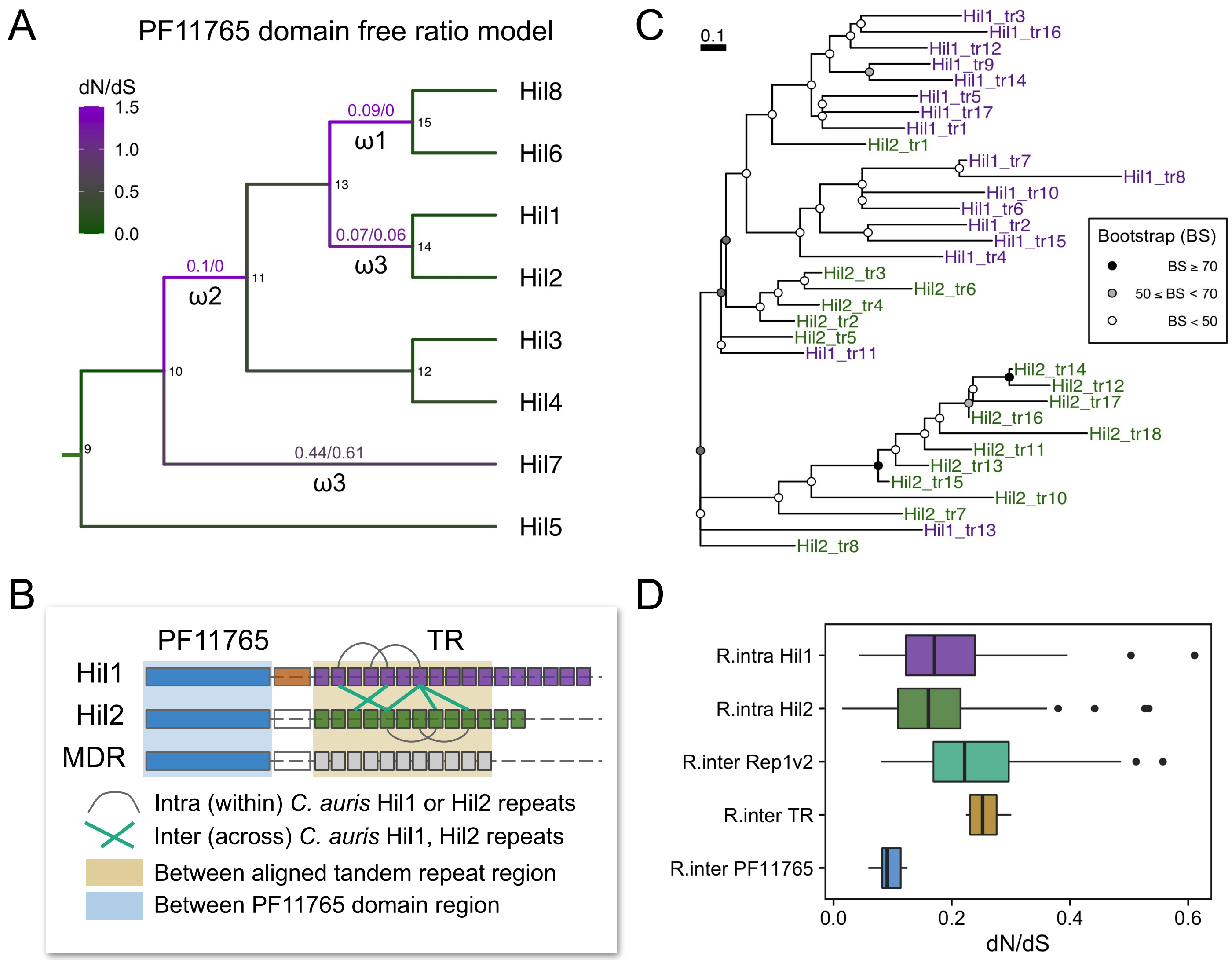


Figure 5. Selective forces on the PF11765 domain in the *C. auris* Hil genes and the expansion of the tandem repeats within Hil1 and Hil2. (A) Phylogenetic tree for Hil1-Hil8 from *C. auris* is based on the PF11765 sequence and shown as a cladogram. Branch colors are based on the estimated dN/dS values. For those with dN/dS > 0.5, the estimates of dN and dS are shown above the branch. The ω1/2/3 below the branches are foreground values used for the branch tests in Table 3. (B) Schematic for the comparisons in D: pairwise dN/dS ratios are estimated between individual 44aa repeats within Hil1 or Hil2 (R.intra Hil1/2, horizontal evolution) or across the two proteins (R.inter Rep1vs2, vertical evolution). An alternative to the "vertical evolution" estimate assumes an aligned portion of the tandem repeat domain is orthologous and pairwise estimates of dN/dS were obtained for *C. auris* Hil1, Hil2 and closely related MDR clade homologs (R.inter TR). For comparison, pairwise dN/dS ratios were also estimated for the PF11765 domain (R.inter PF11765). The orange box between the PF11765 and TR domains indicates a Serine-rich repeat region only present in Hil1. (C) Maximum likelihood tree for the first 17 repeats from Hil1 and Hil2 suggests most of the repeats likely originated after gene duplication. Branch length is in the unit of substitutions per codon. Repeats from Hil1 are in purple and those from Hil2 are in green. White, gray or black circles on the ancestral nodes indicate bootstrap support levels. (D) Pairwise dN/dS ratios estimated using the YN00 program in PAML are shown as boxplots, where the box shows the interquartile range (IQR), the upper and lower whiskers extend to the largest and smallest values no further than 1.5 x IQR, the middle line shows the median and dots show outliers beyond the 1.5 x IQR.

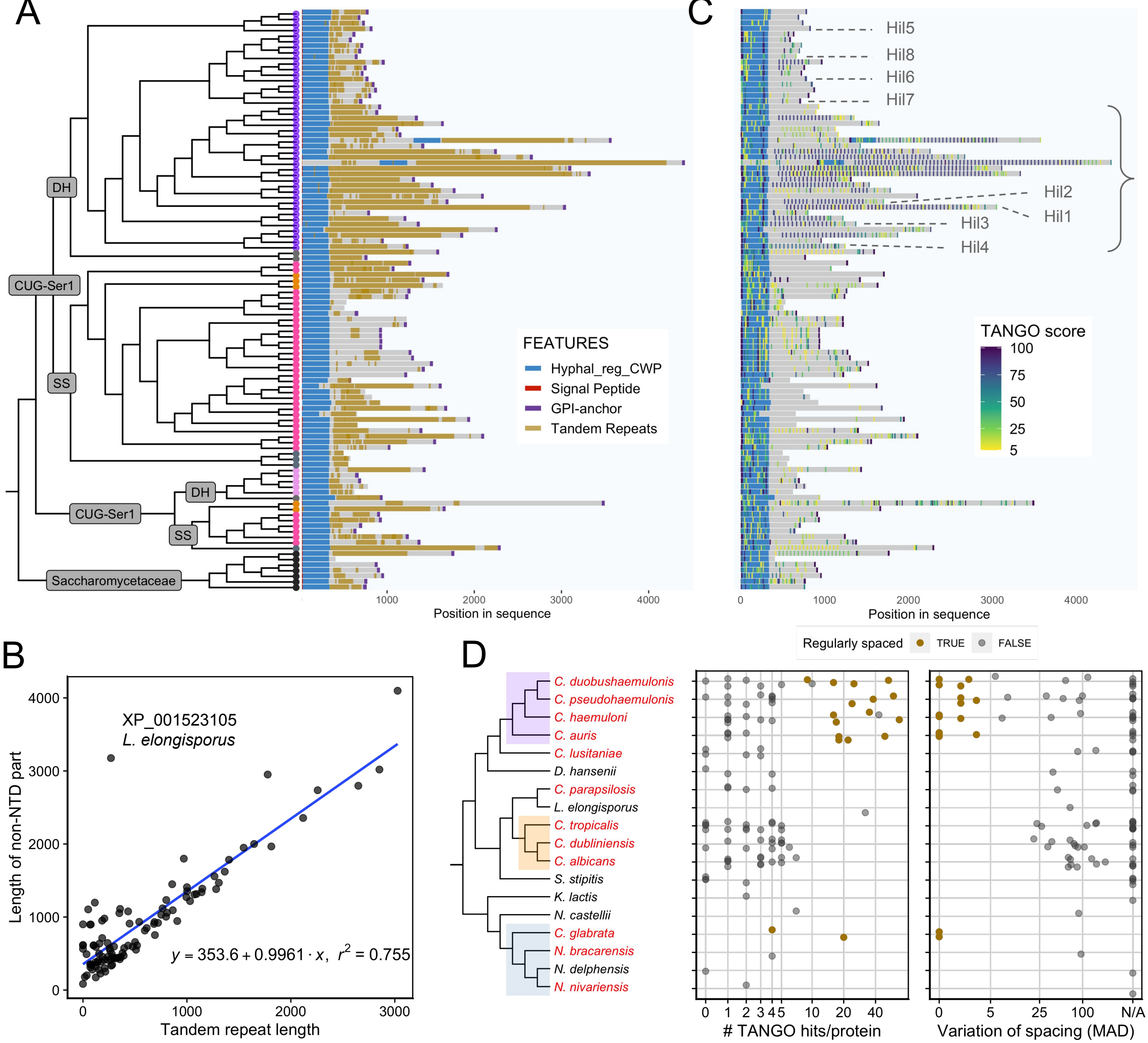


Figure 6. Divergence in the yeast Hil family in length and central domain features. (A) Domain architecture plot showing that the majority of the homologs have a signal peptide and a GPI-anchor at the two termini, with the PF11765 domain at the N-terminus followed by a central domain that is highly repetitive. (B) x-y plot showing length of the non-PF11765 (NTD) portion of a Hil family protein as a function of the length of its tandem repeat sequences. The linear regression line is shown in blue, with parameters and r^2 values below. An outlier to the trend is labeled. (C) Distribution of TANGO predicted β-aggregation sequences. The median per-residue probability is used as the score for each sequence and is shown in a color gradient. A group of MDR clade sequences are labeled by a curly bracket. These sequences uniquely harbor a large number of regularly spaced TANGO hits. The eight *C. auris* Hil genes are labeled. (D) The left panel shows the species tree. The middle panel plots the number of strong TANGO hits (score >= 30) per sequence, grouped by the species, and the right plot shows the variance in their inter-TANGO-hit spacing for the same proteins (MAD = median absolute deviation). Proteins with more than three strong TANGO hits and a MAD of the spacing less than 5 residues are labeled as "regularly spaced" and shown in gold color.

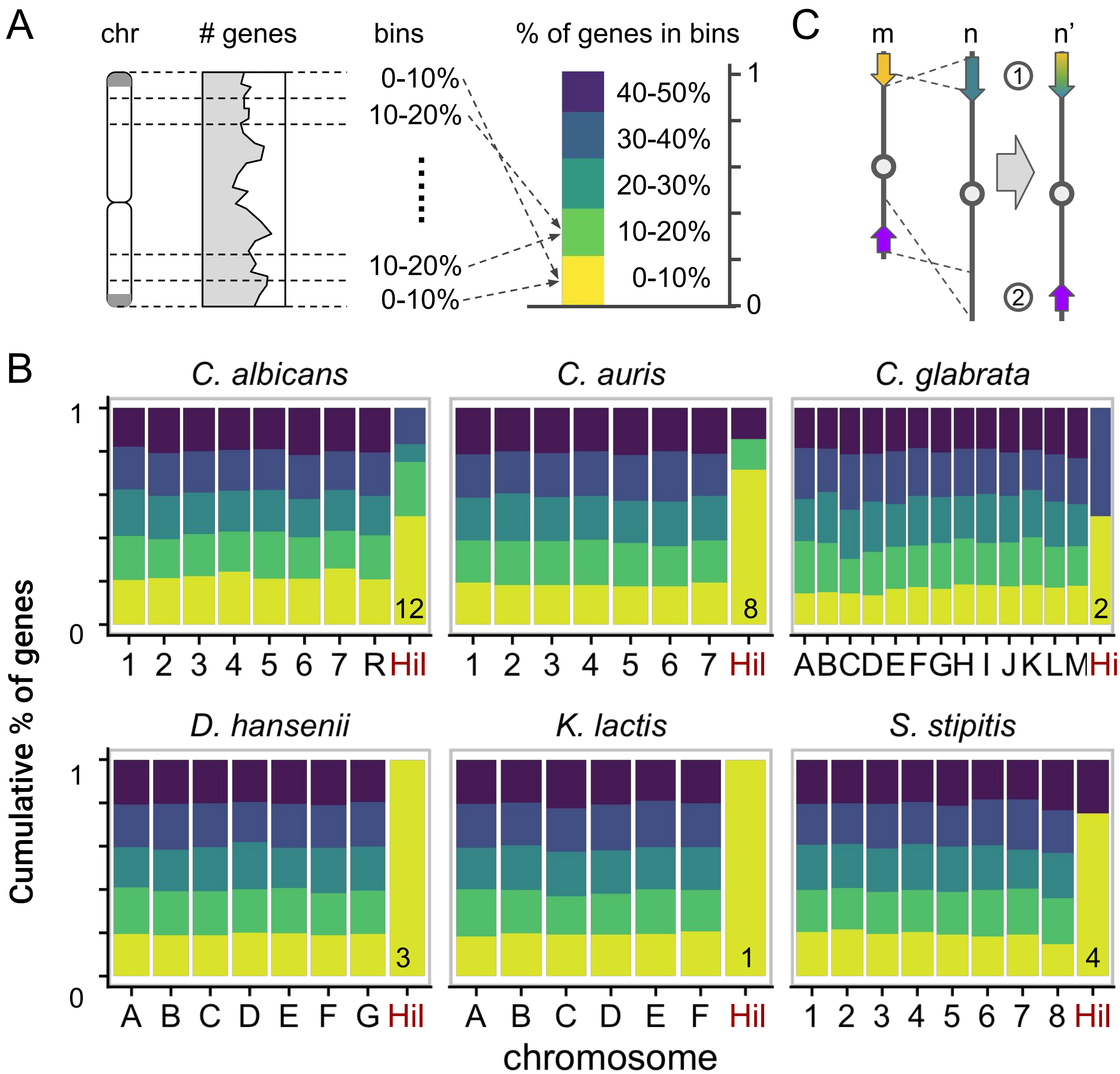
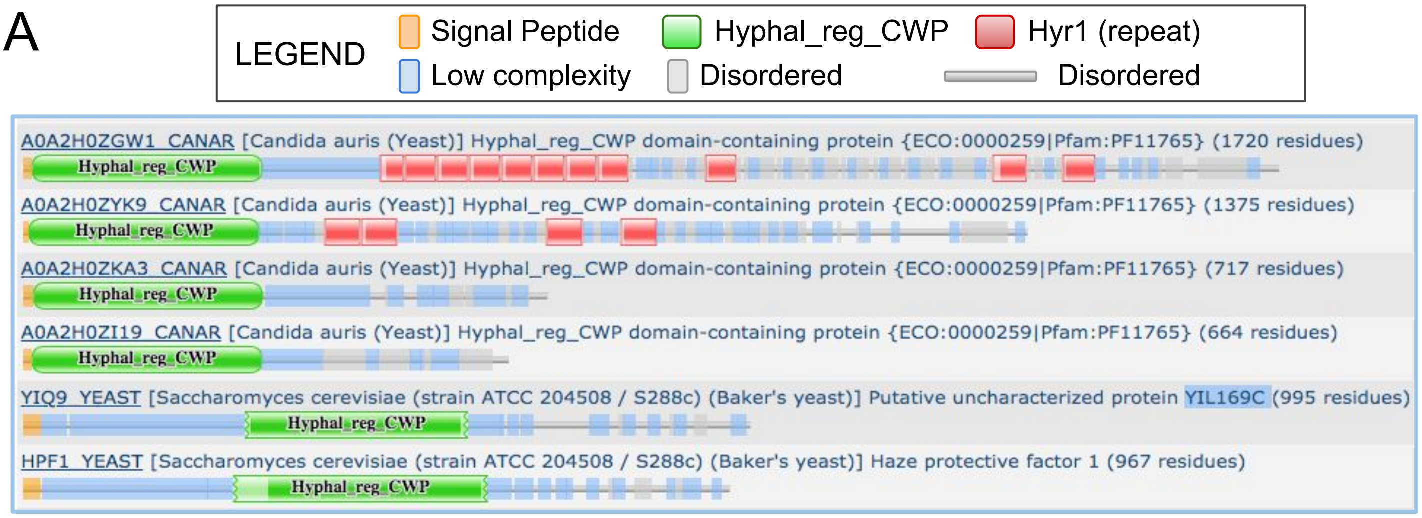


Figure 7. Hil homologs are preferentially located towards the chromosome ends. (A) Schematic of the analysis: each chromosome (chr) is folded and divided into five equal-length "bins" that are ordered by their distance to the nearest telomere (gray). The cumulative bar graph on the right summarizes the distribution of genes along the chromosome. (B) This method is applied to six species with a chromosomal level assembly. The Hil homologs in each species are plotted in their own group with the family size labeled at the bottom. A goodness-of-fit test comparing the distribution of the Hil genes to the genome background yielded a *P*-value of 3.6×10^{-6}. (C) Ectopic recombination between subtelomeres could facilitate (1) creation of a new family member by recombination between two existing members and (2) duplication of a subtelomeric gene onto the equivalent region on a different chromosome.

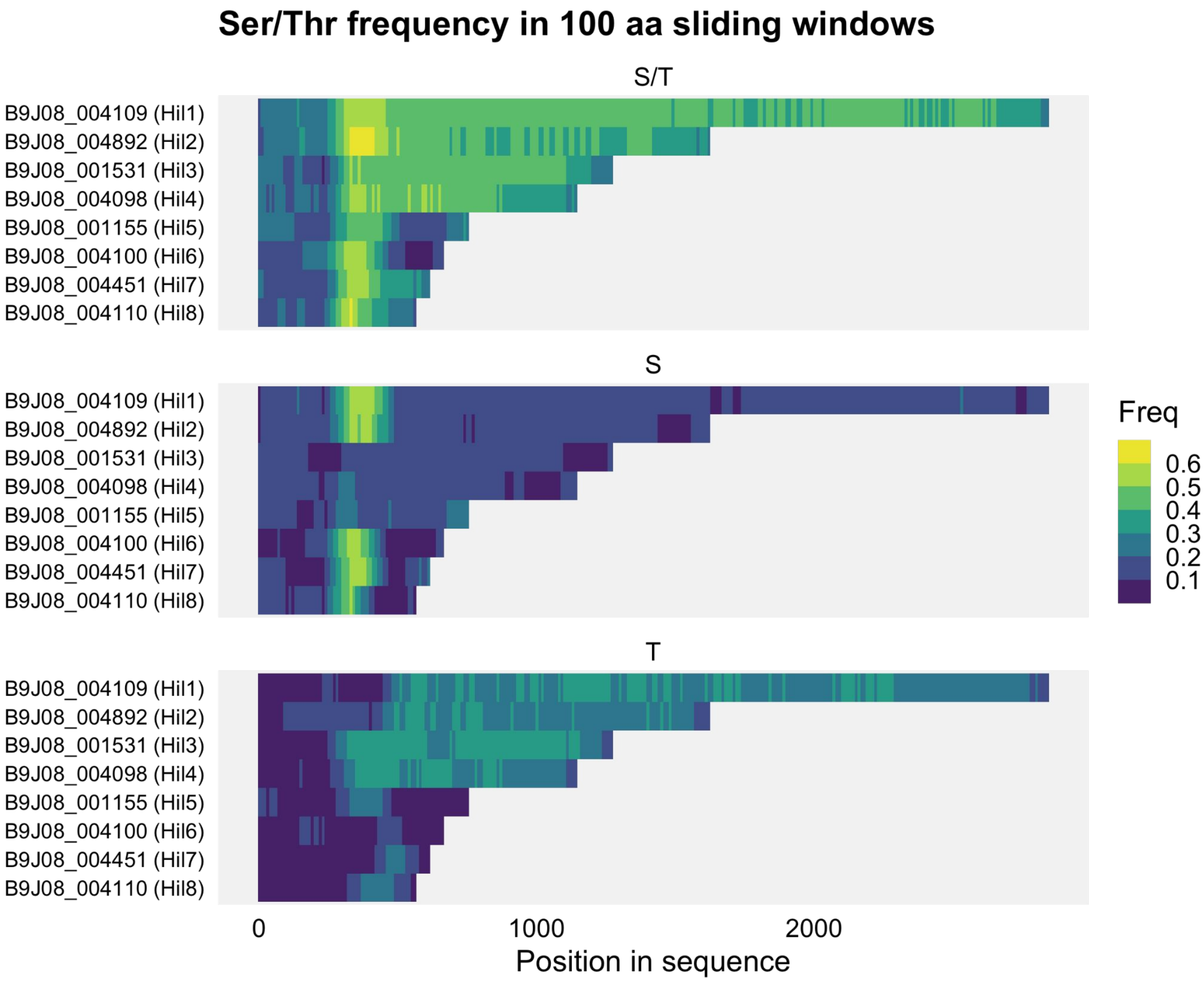
A



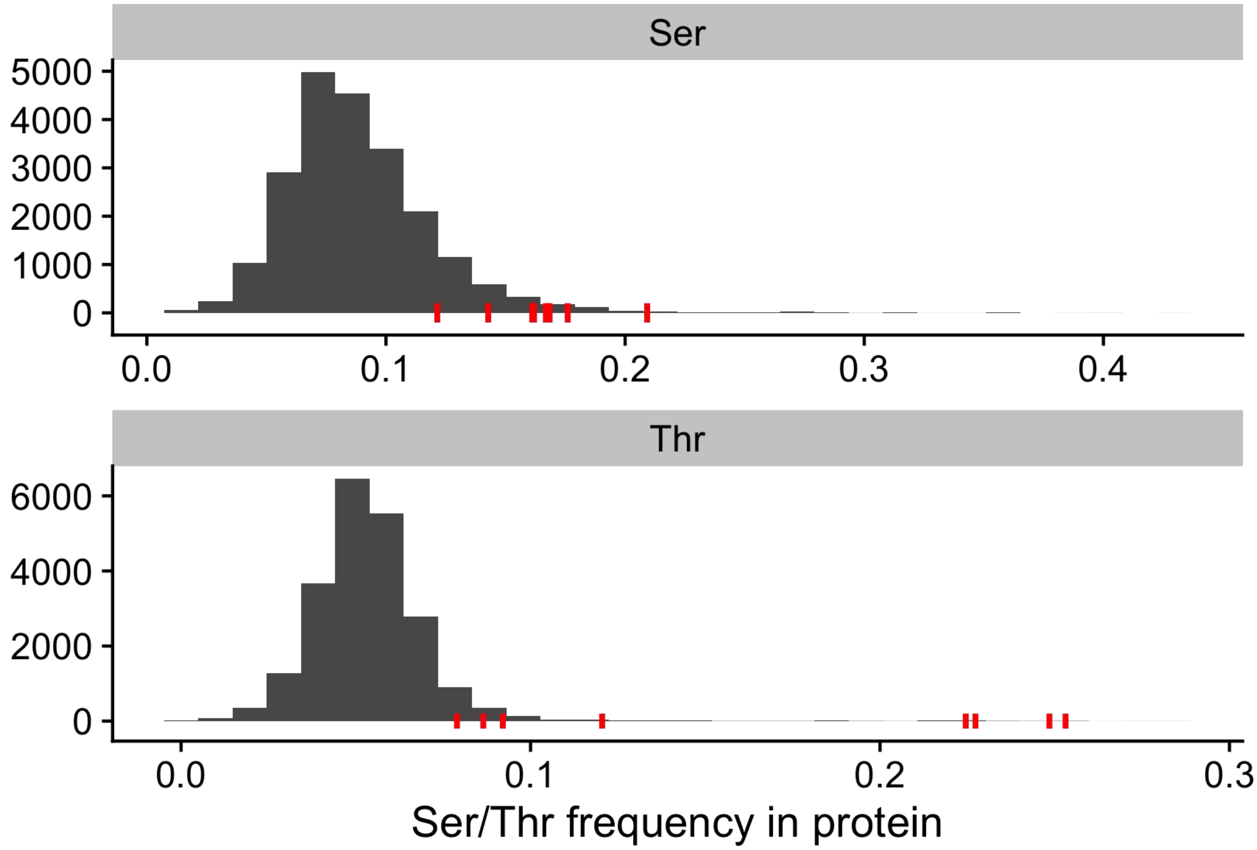
B

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
	CAGL0I07293g_XP_447567.2_Hyphal_reg_CWP	83.2	83.2	91%	4e-22	28.12%	Query_13989
	CAGL0E06600g_XP_445977.1_Hyphal_reg_CWP	76.6	93.2	98%	8e-20	28.31%	Query_13986
	YOL155C_HPF1	42.0	58.5	56%	3e-08	27.66%	Query_13988
	YIL169C_CSS1_Hyp_reg_CWP	39.3	39.3	76%	2e-07	23.78%	Query_13987

Supplementary figure 1. Two *S. cerevisiae* proteins with the PF11765 domain have different architecture and are more divergent from *C. auris* Hil proteins than Hil homologs from the equally distant *C. glabrata*. (A) Comparing the domain architectures of the two *S. cerevisiae* proteins with the PF11765 (Hyphal_reg_CWP) domain to the Hil homologs from *C. auris*. Notice the *S. cerevisiae* proteins are distinct in that their PF11765 domain is in the middle rather than the N-terminus of the protein. (B) BLASTP comparison with *C. auris* Hil1's PF11765 domain as query and the two *C. glabrata* and two *S. cerevisiae* proteins as subjects. *C. glabrata* is in the same family as *S. cerevisiae* and equally distantly related to *C. auris*. Notice the much lower query coverage and less significant E-values for the *S. cerevisiae* sequences.



Supplementary figure 2. Ser/Thr frequency in the *C. auris* Hil family. The Ser+Thr or the individual amino acid frequencies were calculated in 100 aa sliding windows with a step size of 10 aa and plotted as a heatmap.



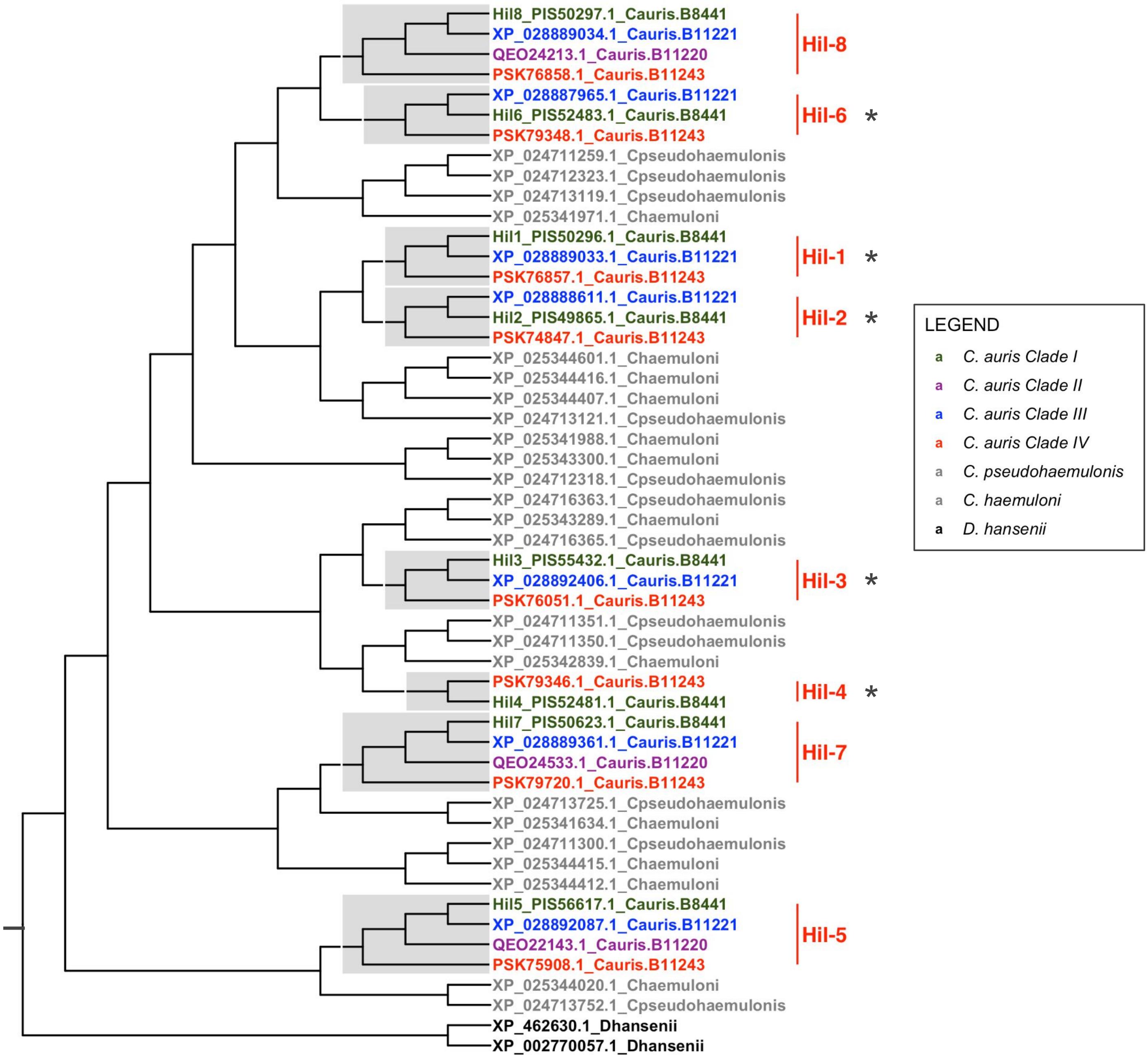
Supplementary figure 3. Comparing the Ser/Thr frequencies in *C. auris* Hil family members with all protein-coding genes in *C. auris*. B8441 strain genome is used for this analysis. The frequency of Ser or Thr residues as a percent of the entire protein is plotted as a histogram for all protein-coding genes. Red ticks indicate the eight Hil genes.

	Hil7	Hil5	Hil3	Hil4	Hil1	Hil2	Hil16	Hil8
Hil7	100	32	36	40	38	41	40	40
Hil5	32	100	41	43	42	44	40	41
Hil3	36	41	100	71	61	62	56	55
Hil4	40	43	71	100	65	67	59	62
Hil1	38	42	61	65	100	83	66	65
Hil2	41	44	62	67	83	100	67	62
Hil6	40	40	56	59	66	67	100	71
Hil8	40	41	55	62	65	62	71	100

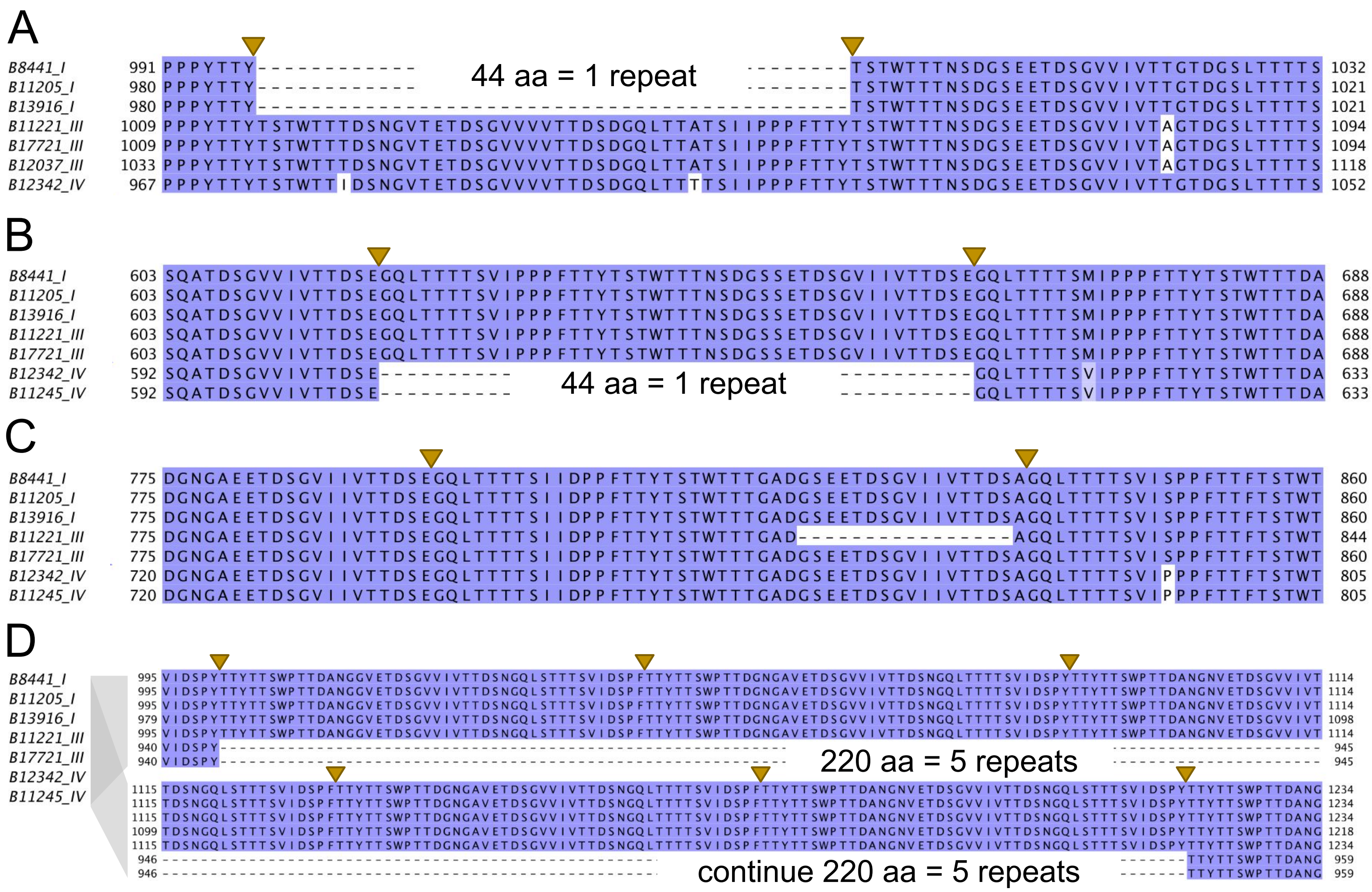
Supplementary figure 4. Percent sequence identity between the PF11765 domains of the eight *C. auris* Hil proteins. Multiple sequence alignment for the eight PF11765 domain sequences were constructed using Clustal Omega and the percent identity matrix reported by the aligner is reproduced as a heatmap (green = low; yellow = medium; red = high).

```
 543- GVETDSGVVIVTTDSDGSLTTTTSVIPPPFTTYTSSWVTTNSAG - 586
 587- ETETDSGVVVVTTNSEGELTTSTSVIPPPYTTYTSTWTTTDGNG - 630
 631- DVETDSGVVIVTTGSDGSLTTTTSVIPPPFTTFTSTWTTTNTDG - 674
 675- ETETDSGIVVVTTDSNGQLTTSTSIIPPPFTTYTSTWTSSQSDG - 718
 719- SEVTDSGVVIVTTDSDGSLTTTTSVIPPPFTTYTSTWATTNSNG - 762
 763- ETETNSGVVVVTTGSDGELTTTTSTIPPPFTTYTSTWISTNSNG - 806
 807- ATETDSGVVVVTTDSDGALTTTTSIIPQPFTTYTSTWTSTNSDG - 850
 851- DTETDSGVVVVTTNSDGALTTSTSVIPQPFTTFTSTWTSSNSNG - 894
 895- AVQTDSGVVIVTTGSDGSLTTTTSVIPPPFTTYTSTWTSSNSDG - 938
 939- ETETDSGVVVVTTDSNGELTTTTSIIPPPFTTFTSTWTSTKSDG - 982
 983- AVETDSGVIIVTTNSEGDLTTTTSIIPPPYTTYTSTWTTTDSNG -1026
1027- VTETDSGVVVVTTDSDGQLTTATSIIPPPFTTYTSTWTTTNSDG -1070
1071- SEETDSGVVIVTAGTDGSLTTTTSVIPPPFTTYTSTWITTNSNG -1114
1115- AVETDSGIIVVTTNSGGSLTTSTSVLPTPFTTYTSTWTTSDGDG -1158
1159- NVQTDSGVVIVTTGSDGALSTTTSVIPPPFTTYTSTWISTNSDG -1202
1203- ETETDSGVVVVTTDSNGALTTTTSIIPPPFTTFTSTWTTTDENG -1246
1247- ATETDSGVVVVTTGTDGSLTTTTSVIPPPYTTFTSTWTTSNSNG -1290
1291- DIETDSGVVIVTTNSDGSLTTTTSVIPPPYTTFTTTWATTNSDG -1334
1335- TTETDSGVVIVTTDSEGQLTTTTSVIPPPFTTYTSTWTSNKSDG -1378
1379- AVETDSGVVIVTTDSDGALTTTTSIIPQPFTTYTSTWTSTNSNG -1422
1423- AIETESGVVVVTTDSNGALTTSTSVIPLPLTTFTTTWTTTNSAG -1466
1467- ETETDSGVVVVETNSNGALTTTTSTFPEPFTTFTSTWTTTDDSG -1510
1511- AIATDSGVVIVTTGSDGSLSTTTSVIPPPFTTYTTTWTSTNSNG -1554
1555- GVETDSGVVIVTTNSDGALETTTSVIDPPFNTYTSTWTTTDADG -1598
1599- AIETDSGVVVVTTGSDGSLTTTTSVIPHPFTTYTSTWVTTGSDG -1642
1643- DTETDSGVIVVTTDSDGALTTSTSLLPVPFTTYTSTWTITNSDG -1686
1687- SQATDSGVVIVTTDSEGQLTTTTSVIPPPFTTYTSTWTTTGANG -1730
1731- GEETDSGVIIVTTDSDGQLATTTSVIPPPFTTFTSTWTTTNSDG -1774
1775- NQATDSGVVIVTTDSDGQLTTTTSVIPPPFTTYTSTWTTTDGNG -1818
1819- AEETDSGVIIVTTDSEGQLTTTTSVIPPPFTTYTSTWTTTGADG -1862
1863- SEETDSGVIIVTTDSAGQLTTTTSVIPPPFTTFTSTWTTTDGNG -1906
```

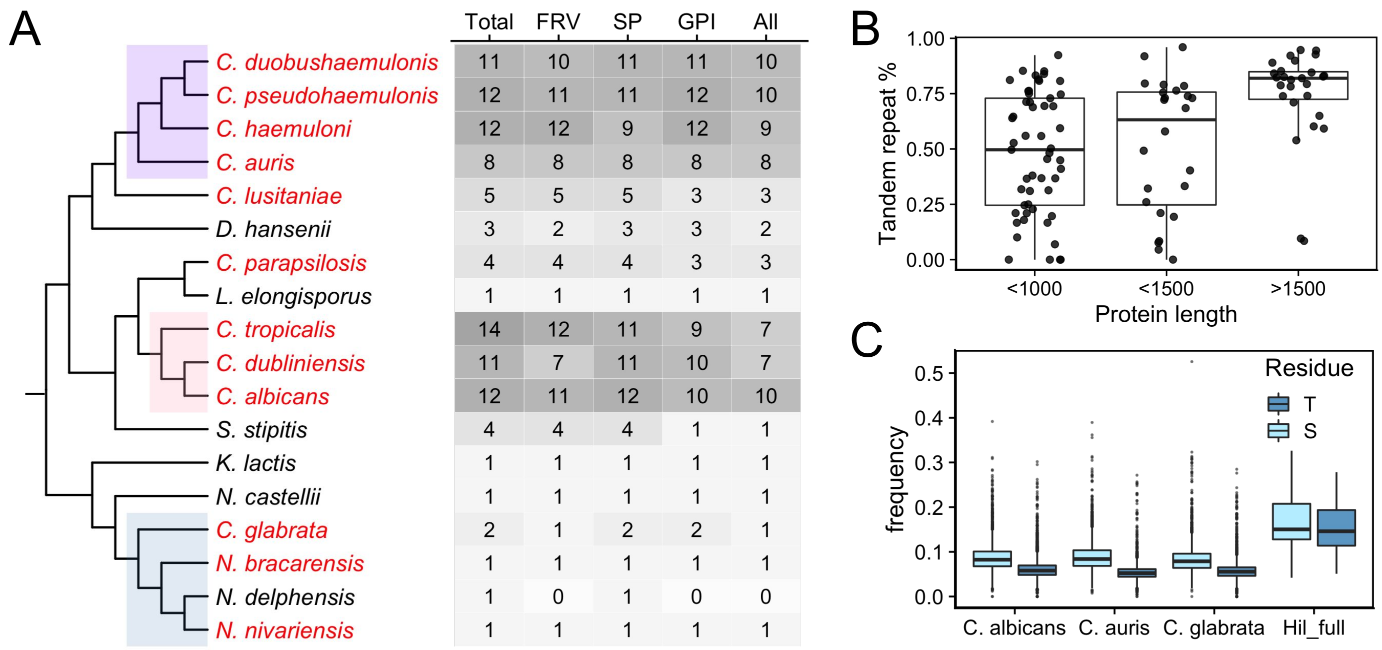
Supplementary figure 5. Tandem repeats in the *C. auris* Hil1 central domain. 31 of ~50 tandem repeat copies are shown with a conserved 44 aa period. The remaining copies show similar patterns but are less conserved in length and sequences. Yellow highlights show predicted β-strands by PSIPred; magenta and plum fonts indicate sequences predicted by TANGO to have strong (>90%) or moderate (30-90%) β-aggregation potentials. WebLogo (above) for the pseudo-alignment of the repeats is created by weblogo.berkeley.edu/logo.cgi



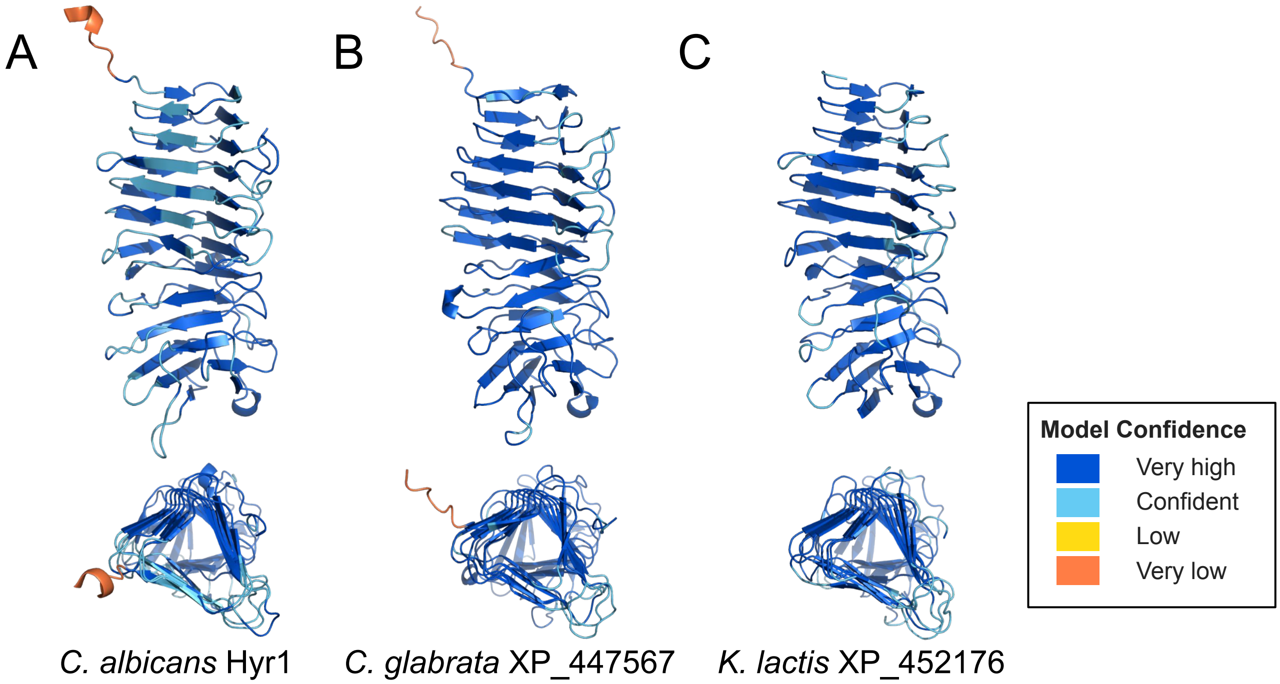
Supplementary figure 6. Reconciled PF11765 domain tree for the Hil family genes in the four clades of *C. auris* strains and two closely related species. The tree is rooted by the two homologs from the outgroup *D. hansenii*. The domain tree was reconciled with the species/strain tree based on (Muñoz 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.



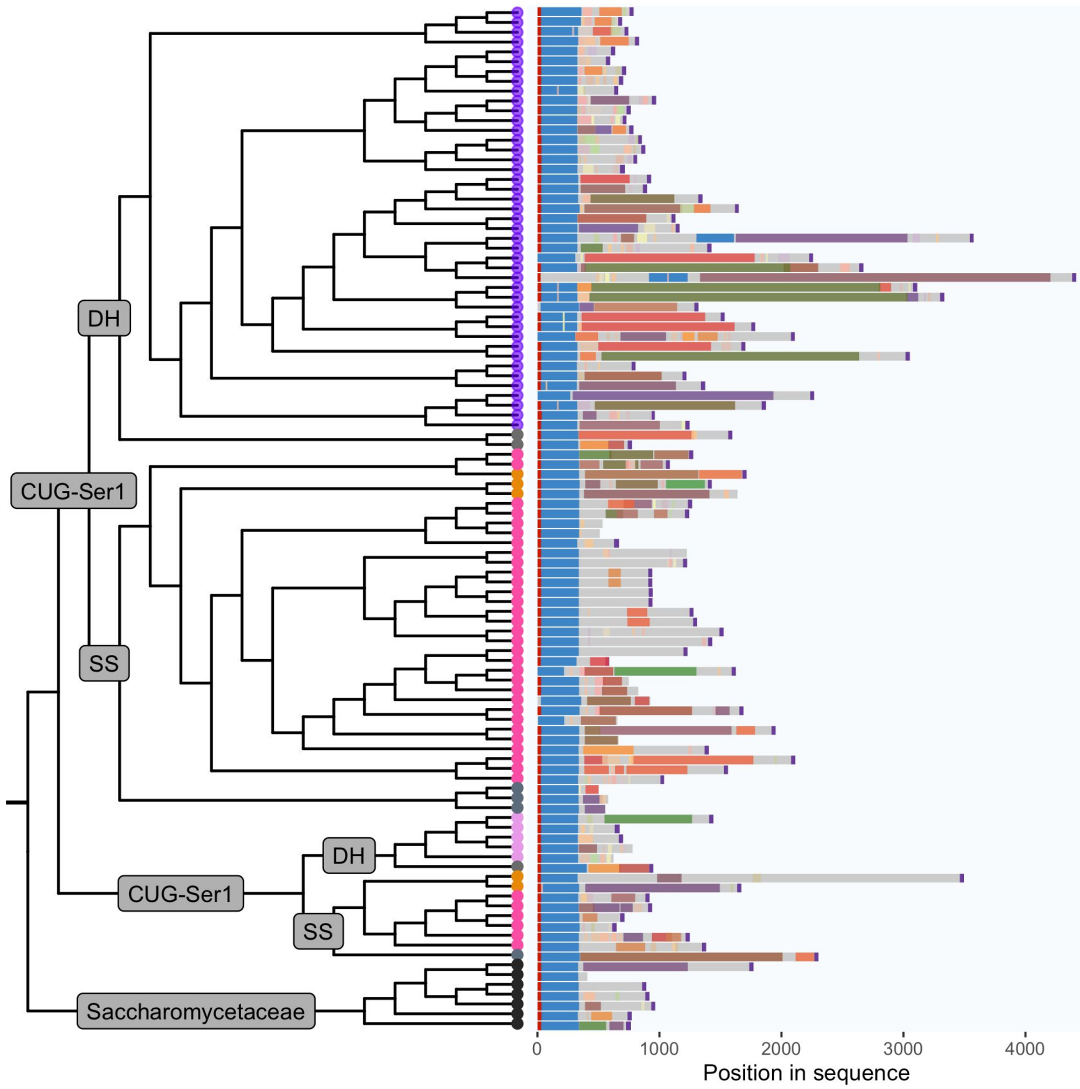
Supplementary figure 7. Examples of tandem repeat copy number variation in Hil1-Hil4 among the *C. auris* strains. (A) A 44 aa indel in Hil1 removes exactly one repeat in all three Clade I strain orthologs. (B) A similar indel polymorphism of exactly one repeat length in Hil2 affecting the Clade IV strains. (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.



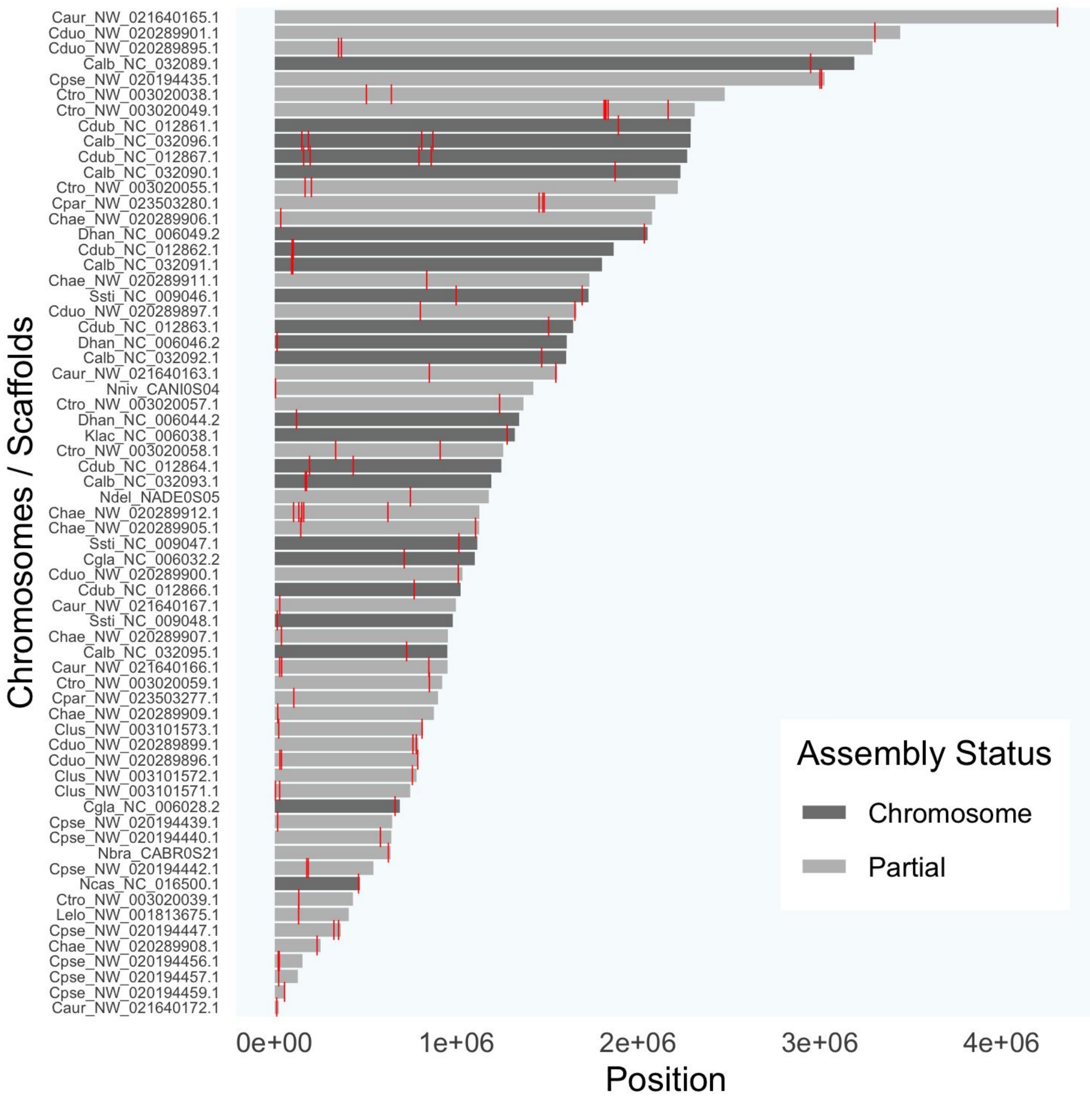
Supplementary figure 8. Majority of the yeast Hil family genes are likely to encode adhesins. (A) Species tree with a table showing the total number of Hil family genes and the subset that pass one of the three tests separately and together (All). The three tests are: positive prediction by FungalRV (FRV), signal peptide prediction by SignalP (SP) and GPI-anchor prediction by PredGPI (GPI). (B) Boxplot for the proportion of a protein identified as tandem repeats, excluding the PF11765 domain. The Hil family genes are divided into three groups based on the full protein length. The box shows the interquartile range (IQR); the upper whisker extends to the largest value no further than 1.5 x IQR and similarly for the lower whisker; the middle lines shows the median. Individual proteins are plotted as dots, with their x-values slightly shifted to avoid overplotting. (C) Genome-wide distribution of Thr/Ser frequencies in the entire protein in three species, compared with that in all Hil proteins (Hil_full). The box plot features are the same as in B except in this case the dots represent outliers beyond the 1.5 x IQR.



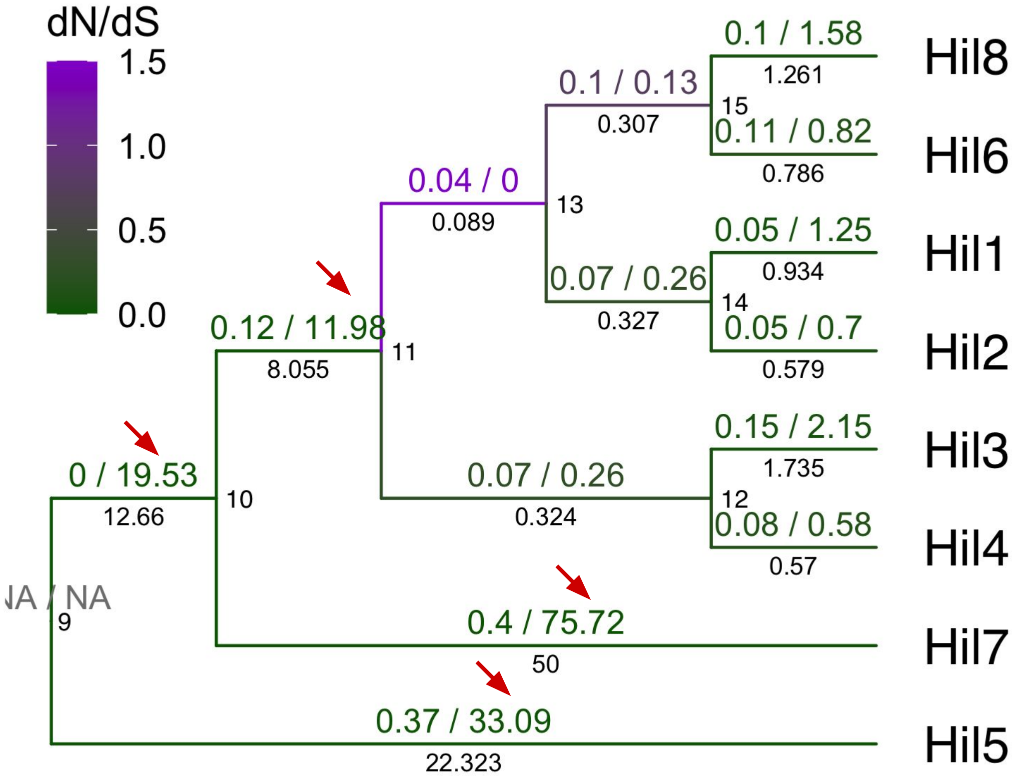
Supplementary figure 9. AlphaFold2 predicted structures for the PF11765 domain in three distantly related Hil homologs. The predicted structures are aligned in PyMol and presented in either longitudinal (top) or cross sectional (bottom) view, highlighting the similarities among the three structures made of repeating β-strands forming a superhelix. Panels A, B and C correspond to three Hil proteins from distantly related species as indicated below the cross-sectional view.



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 in the current figure tandem repeats belonging to different sequence clusters as determined by XSTREAM are shown in different colors.



Supplemental figure 11. Chromosomal locations of the Hil family genes. Each row is either an assembled chromosome (dark grey) or a scaffold (light grey). The length of the bar corresponds to the length of the chromosome or scaffold, whose NCBI IDs are listed on the left. The location of the Hil genes are labeled as red vertical stripes.



Supplemental figure 12. **dN/dS estimates for the PF11765 domain in the *C. auris* Hil family.** Same as Figure 5A except a F3x4 model ("CodonFreq = 2") instead of F1x4 ("CodonFreq = 1") was used to estimate the codon frequencies. Also, dN and dS values are labeled on top of all branches to show the unusually high dS estimates on some of them (red arrows). The branch length, defined as the estimated number of substitutions per codon, is labeled under each branch.