Text S1 - Identify Hil family homologs

Accurately and comprehensively identifying Hil family homologs is the basis for all downstream analyses. Therefore, we detail the multiple steps we took to minimize biases and errors. Despite our best efforts, our conclusions are only as good as the quality of the genome assemblies, some of which need further improvement as revealed by our analyses. The use of long-read technologies such as Oxford Nanopore and PacBio in combination with Illumina sequencing shows great promise. However, careful bioinformatic processing, including assembly and manual curation, is also critical to achieving accurate homolog sequences.

## First round of identification using a single PF11765 domain as the query

To identify all Hil family homologs, we first used the *C. auris* Hil1 Hyphal\_reg\_CWP domain to query the RefSeq database (May 2022). All 189 BLASTP hits that passed the *E*-value threshold of 10-5, with a query coverage greater than 50%, were from Ascomycota (yeasts); all but one was from the Saccharomycetes class (budding yeast). A single hit was found in the fission yeast *Schizosacchromyces cryophilus*. We used that hit as the query and searched against all fission yeasts in the nr protein database with a relaxed E-value cutoff of 10-3, and identified no further hits. We therefore conclude that the Hil family, defined by the presence of the Hyphal\_reg\_CWP domain, is yeast specific.

## Expanded search using multiple Hil homologs’ PF11765 domain as queries

We next expanded the search with additional queries from two Hil family homologs, one from *C. albicans* (Hyr1) and another from *C. glabrata* (XP\_445977). Based on an initial gene tree for all homologs identified in the first round, the three queries belonged to different clades spanning the entire tree. We reasoned that adding the *C. albicans* and *C. glabrata* queries would avoid biasing the search towards homologs more closely related to our initial *C. auris* query protein. The new search results added just three hits to our first-round list, suggesting that our initial list included most of the homologous proteins.

## Including additional yeast genome resources

The Genome Resources for Yeast Chromosomes (GRYC, <http://gryc.inra.fr/>) includes additional yeast genomes not present in the NCBI RefSeq database, e.g., the Nakaseomyces genus that includes *Candida* pathogens closely related to *C. glabrata*. We performed the same search in this database and recovered 16 additional homologs belonging to seven species.

## Excluding species from the analysis

To infer the evolutionary history of the Hil family, including the duplication events, we relied on the recently published phylogeny for the budding yeast subphylum with 332 species (Shen *et al.* 2018). We manually added *C. duobushaemulonis*, *C. pseudohaemulonii* and *C. haemuloni* to the species tree based on a separate phylogeny (Muñoz *et al.* 2018). Among the 35 species that had at least one Hil homolog as identified above, three were excluded because they were not part of the species tree. These include *Diutina rugosa*, *Kazachstania barnettii* and *Artibeus jamaicensis*. The remaining 32 species represent a broad sampling of the budding yeast subphylum. The final list of Hil family homologs includes 215 genes.

## Verify RefSeq protein sequence annotation using newer assemblies

The RefSeq database provides a set of non-redundant and well-annotated sequences that can serve as a stable reference for gene identification and characterization. Because it emphasizes stability, newly sequenced and assembled genomes using more advanced technologies often take time to be integrated into the database. As a result, some of the RefSeq sequences may contain annotation errors. In a TBLASTN search using the RefSeq hits in *C. tropicalis* against a long-read based assembly for the same MYA-3404 strain (assembly ID: GCA\_013177555.1), we found that several homologs had a more downstream stop codon in the new assembly. We used the ORF finder (<https://www.ncbi.nlm.nih.gov/orffinder/>) to identify the longest ORF in those cases. Of the 16 homologs identified in this species, 6 were substantially longer in the new assembly than in the RefSeq assembly, and 2 were predicted to be substantially shorter (Table S8). It is worth noting that while the new assembly is likely to be generally more reliable, further experimental validation will be needed to determine the sequences for the genes in conflict. To determine if this discrepancy represents a more wide-spread problem, we repeated the same search for species that have a long-read based assembly available, albeit not for the same strain as the RefSeq assembly, including *S. stipitis* (NRRL Y-7124, GCA\_016859295.1), *C. albicans* (NCYC4166, GCA\_005890745.1), *C. parapsilosis* (CBS6318, GCA\_00982555.2) and *C. glabrata* (BG2, GCA\_014217725.1). For *S. stipitis*, 16/18 RefSeq hits were shorter than 600 aa, and 15 were labeled as “incomplete”. We identified seven hits with a query coverage above 50% in the new assembly, all of which are longer than 900 a.a. For the other three genome assemblies, the RefSeq hits were highly consistent in sequence with the new assembly, despite biological differences between strains. We conclude that the issues with the *C. tropicalis* and *S. stipitis* RefSeq hits are specific to their assemblies and not a general issue with the RefSeq database. More importantly, the sequence of the PF11765 domain was always highly consistent between the RefSeq and the new assemblies, even for *C. tropicalis*. As our phylogenetic reconstruction is based solely on the alignments of the PF11765 domain, assembly quality issues would have had no impact on the inference of the evolutionary history for the Hil family. Nonetheless, efforts to improve the genome assembly and update the RefSeq database are crucial for future studies to characterize gene family evolution, particularly when the gene family is repeat-rich.

## Identification of additional Hil family homologs in *C. glabrata* using the PacBio assembly

The recent sequencing and assembly of the *C. glabrata* reference strain using PacBio (Xu *et al.* 2020) revealed that the subtelomeric regions in the RefSeq assembly (GCF\_000002545.3) were incomplete. To determine if our initial homology search missed any Hil family homologs in *C. glabrata*, we repeated the search in the new assembly (GCA\_010111755.1). Of the 13 hits that passed all criteria, only three were identified in the RefSeq assembly. 12 of the 13 hits were located in the subtelomeric regions, including all 10 that were identified in the new assembly alone. Since many yeast adhesin families, including the Hil family, are enriched in subtelomeric regions, we repeated this analysis for additional species with a long-read assembly available, to determine how widespread the issue of missing homologs as observed in *C. glabrata* is. We focused on the genomes assembled at least to a chromosomal level, including *S. cerevisiae* (GCA\_016858165.1), *K. lactis* (GCA\_007993695.1), *C. nivariensis* (GCA\_017309295.1) and *C. albicans* (GCA\_005890745.1). In *S. cerevisiae* and *K. lactis*, the same number of hits (0 and 1) were identified in the long-read and RefSeq assemblies. In *C. nivariensis*, we identified three hits in the long-read assembly compared with two in the RefSeq assembly. Finally, in *C. albicans*, we identified 13 hits in the long-read assembly compared with 12 hits in the RefSeq assembly. However, two of the 13 hits were identical in the nucleotide sequence, raising questions as to whether they resulted from recent duplications or were due to assembly errors. Although we only tested four more species, limited by the availability of long-read based assemblies, these results suggest that the large discrepancy in the Hil family size observed in *C. glabrata* is unique and may have to do with the special challenges in assembling its subtelomeric regions. Nonetheless, we believe that improved assemblies for all species are urgently needed for studying adhesin families.

### Literature cited in Text S1

Muñoz J. F., L. Gade, N. A. Chow, V. N. Loparev, P. Juieng, *et al.*, 2018 Genomic insights into multidrug-resistance, mating and virulence in Candida auris and related emerging species. Nat. Commun. 9: 5346. https://doi.org/10.1038/s41467-018-07779-6

Shen X.-X., D. A. Opulente, J. Kominek, X. Zhou, J. L. Steenwyk, *et al.*, 2018 Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. Cell 175: 1533–1545. https://doi.org/10.1016/j.cell.2018.10.023

Xu Z., B. Green, N. Benoit, M. Schatz, S. Wheelan, *et al.*, 2020 De novo genome assembly of Candida glabrata reveals cell wall protein complement and structure of dispersed tandem repeat arrays. Mol. Microbiol. 113: 1209–1224. https://doi.org/10.1111/mmi.14488