Title: Defining the *Candidozyma auris* pan-genome and essentiality

Authors: Joseph J Hale*[1], Ajay J Larkin*[1], Jackson R Rapala[1,2], Rebecca Hurto[3], Guolei Zhao[1], Brynn E Elson[4], Lydia Freddolino[2,3], Evan S. Snitkin*[1], Teresa R O'Meara*[1]

Affiliations:

[1]Department of Microbiology and Immunology, University of Michigan Medical School; Ann Arbor MI 48019 USA.

[2]Gilbert S. Omenn Department of Computational Medicine and Bioinformatics, University of Michigan Medical School; Ann Arbor MI 48019 USA.

[3]Department of Biological Chemistry, University of Michigan Medical School; Ann Arbor MI 48019 USA.

[4]Program in Immunology, University of Michigan Medical School; Ann Arbor MI 48019 USA.

* denotes equal contribution

Correspondence to:

Teresa O'Meara (tromeara@umich.edu)

Evan Snitkin (esnitkin@umich.edu)

Abstract:

Candidozyma auris is an emerging multi-drug resistant fungal pathogen characterized by high mortality and rapid transmission in healthcare settings, but the genetic drivers of phenotypic variation between strains and the landscape of gene essentiality in this organism remain undercharacterized. Here, we integrate pangenomic analysis with global essentiality screening to establish a foundational understanding of the *C. auris* genome and identify potential therapeutic targets. We performed pangenome analysis on 695 outbreak strains of *C. auris* selected to be genetically representative of publicly sequenced genomes. After using BLAST to refine the pangenome, we found that 96.8% of gene families were core, with the remaining high-confidence accessory gene families primarily consisting of gene loss events or clade-specific genes. The high proportion of core genes emphasizes the clonal nature of these outbreak strains, but comparative analysis with the closely related *C. haemuli* species complex suggested that most of these core genes are functionally dispensible. To examine this hypothesis, we developed a novel insertional

mutagenesis approach that leverages the promiscuous integration of linear DNA in the *C. auris* genome. This global analysis identified 614 high-confidence essential genes. Crucially, nearly one-third of these genes, including the conserved translation initiation factor Sui1, exhibit divergent essentiality patterns compared to the model yeasts *Candida albicans* and *Saccharomyces cerevisiae*. These findings highlight organism-specific biology that would be overlooked by orthology alone. By combining pangenomic diversity with functional essentiality, this study provides a comprehensive resource for identifying species-specific determinants of virulence and prioritizing novel targets for antifungal drug development.

Introduction:

Candidozyma auris is an often multi-drug resistant fungal pathogen that causes hospital-associated outbreaks [1]. The mortality of systemic infections approaches 50%, and cases are on the rise due to the high colonization and transmission rates in healthcare settings. Once systemic disease is established, there are often few options for treatment, as many isolates have developed resistance to at least one of the three commonly used classes of antifungals and pan-resistant isolates have also been observed [2]. However, as a recently discovered organism, *C. auris* has a largely undercharacterized genome, an unknown set of essential genes, and poorly-understood genetic mechanisms for generating phenotypic variation. This lack of knowledge of essential pathways can limit our ability to identify potential new drug targets specific to this species. Additionally, our limited insight into the genetic changes experienced by lineages of *C. auris* can hinder our ability to predict the genotypic and phenotypic adaptations that occur in healthcare settings. Thus, there is a need to better understand the genetic variation that exists within and between the different clades of *C. auris*, as well as investigate the genes and pathways essential for its survival.

Pangenome analysis is an approach to identify the full collection of orthologous gene clusters across all the sequenced genomes for a given species, thus allowing an analysis of genetic diversity that cannot be captured in a single reference strain. Additionally, genes that are present in every isolate of an organism (core genes) are more likely to be essential than those that are variably present between isolates (accessory genes). Several features of *C. auris* make it potentially suitable for pangenome analysis. Previous studies have shown that *C. auris* exhibits high rates of genomic rearrangement, producing significant differences in gene content between isolates [3–5]. Additionally, *C. auris* strains vary within and between clades in several clinically relevant phenotypes, such as antifungal resistance, adherence to skin and abiotic surfaces, and nutrient utilization [1,6–9]. Finally, hospital pathogen surveillance has produced a large number of sequenced isolates from diverse clades, time points, and geographic locations [10–13]. A pangenome analysis of *C. auris* that takes its clonal nature and population structure into account has the potential to reveal the genetic changes experienced by hospital-associated isolates and lineages, in addition to identifying significant and clade-defining genetic variants for future studies.

Beyond presence and absence, essential genes are defined as the set of genes required for viability in non-stress conditions. One approach for defining the essential gene set is to perform insertional mutagenesis, and to use sequencing approaches to identify regions of the genome that do not tolerate insertions. By identifying essential genes, we can gain insight into the underlying biology of an organism, identify candidates for drug targeting [14–19], and uncover species-specific phenotypes [20,21] . In *C. auris,* the *piggyBac* transposon system was used to identify lncRNA that

controls filamentation and *CDT1*, a calcium-calcineurin signalling pathway-sensitive ATPase that controls fluconazole tolerance [22,23]. Additionally, we developed and used insertional mutagenesis through *Agrobacterium*-mediated transformation to identify regulators of filamentation and adhesion [7,24]. However, genome-scale analyses of essentiality in *C. auris* have not been previously described, due to difficulties in achieving saturating mutagenesis and mapping insertion sites at scale. Combining essentiality analysis with pangenomics would allow for further characterization of genes that are both conserved and essential and thus ideal as future drug development targets.

Here, we combine pangenomic analysis with global essentiality screening to build a foundational picture of the *C. auris* genome. We performed *de novo* genome assemblies for a genetically representative set of 695 genomes primarily drawn from outbreak isolates in clades I, III, and IV and used them to define the pangenome. We used ancestral reconstruction to identify accessory gene families with clear signatures of convergent evolution, including the repeated loss of a specific subtelomeric deletion that overlaps a known adhesin gene. This analysis provides a detailed description of accessory gene families at both the species and clade level, in addition to offering insight into the selective pressures experienced by this population. However, we also found that 96.8% of all gene families were categorized as core, appearing in at least 95% of all isolate genomes. The high percentage of core genes prevented us from gaining insight into gene essentiality just from the pangenome. Therefore, to define the *C. auris* conserved and essential genes, we took advantage of the discovery that *C. auris* will promiscuously insert linear DNA into its genome to develop an insertional mutagenesis library and sequencing approach. We identified 614 high confidence predicted essential genes, with nearly a third showing a divergent pattern of essentiality from the model budding yeasts *Candida albicans* and *Saccharomyces cerevisiae*. As a case study, we demonstrate variability in requirement for the Sui1 translation initiation factor between *C. auris* and *C. albicans.* Together, this study provides a comprehensive characterization of the genetic variation that exists within *C. auris* through multiple parallel approaches, building a foundation for future work on identifying novel targets for antifungal therapy or species-specific determinants of virulence.

Results:

Generating a high-confidence *C. auris* pangenome:

To provide a comprehensive view of the currently sequenced *Candidozyma auris* outbreak strains, we first selected a panel of 695 genetically diverse isolates from among the short-read Illumina sequences uploaded to the NCBI pathogens database, based on published SNP trees. For each tree, we removed isolates until all members of the tree differed by at least 30 SNPs from

their nearest neighbor. We also collected and sequenced isolates from the University of Michigan hospital, Rush University Medical Center, and the Michigan Department of Health and Human Services, performing both long and short read sequencing for use in hybrid assembly (**Supplemental Table S1**).

All selected isolates were assembled and annotated via a custom-built snakemake pipeline[25] designed for *C. auris* (Methods) and built on the framework of the annotation tool Funannotate, which was designed for fungal genomes [26]. To ensure potential structural variants and genomic rearrangements were not removed by reference-based alignment, we performed *de novo* assembly for all genomes. For the 41 isolates where we performed paired Illumina and long-read nanopore sequencing, we generated hybrid assemblies, providing several high-quality, chromosome-level assemblies for each of the three clades (**Figure 1A, left panel**). Assemblies were only added to the final dataset if they had nucleotide BUSCO scores above 95%, indicating completeness of the assembly and gene annotation steps (**Figure 1A, right panel**). We also removed a total of six genomes that exhibited signs of contamination.

Given that *C. auris* exhibits several features more commonly associated with bacterial populations, such as a lack of sexual reproduction and highly clonal outbreaks, we performed pangenome analysis and orthogroup clustering with both Panaroo [27] and Orthofinder [28]. Orthofinder is a pangenome analysis tool commonly applied to eukaryotic populations[28]. Alternatively, Panaroo is well-suited to a highly clonal dataset of short-read assemblies where the accessory genome is likely to be small, as its gene refinding algorithms can reduce the impact of assembly and annotation errors that would otherwise introduce spurious gene loss events. Comparing Panaroo outputs to Orthofinder revealed that Panaroo reduced the total size of the pangenome by nearly 35%, while assigning more core gene families in total (5527 vs. 5370). The majority of this difference was attributable to the number of singleton and near-singleton gene families, which was higher in Orthofinder by a factor of over 10X (**Figure 1B**). As this matched our expectations for an inflated number of rare gene families produced by partially-fragmented short-read assemblies, we used the Panaroo results for all further downstream analysis. This resulted in an initial pangenome size of 6183 gene families, 89.4% of which consisted of core genes that appear at least 95% of all isolate genomes, and produced a network graph visualizing differences in synteny and gene order across isolates (**Supplemental Figure 1**).

Even after the reductions to the pangenome size made by Panaroo, we still found that many gene families exhibited signatures of technical artifacts, which we then resolved. For example, the gene adjacency and synteny from Panaroo's pangenome graph revealed that some gene families appeared to be isolated on the graph with no connecting edges, indicating that these genes were never found on the same contig or chromosome as any other gene family in any

isolate (**Supplemental Figure 1**). Additionally, direct inspection of our assemblies revealed an apparent overreporting of gene absences, where a gene family would only be annotated and marked as present in a fraction of the genomes it appeared in. While these technical issues are not unexpected and have been noted in previous publications [29], they still presented a significant obstacle to our analysis, which primarily focused on accessory rather than core gene families. To verify that the accessory gene families present in our dataset were accurate and biologically meaningful, we designed a pipeline to validate the gene presence and absence data reported by pangenome tools using BLAST+ [30]. This tool takes a specific gene family as input and performs a BLAST search against every isolate in the pangenome, regardless of whether the gene was originally reported as present or absent. Highly stringent BLAST search parameters (>90% sequence identity, >90% coverage, and e-value $<10^{-5}$) and gene adjacency information from the pangenome graph minimize the chance for off-target hits beyond the target gene family (Methods), allowing us to identify cases of overreported gene absence with high confidence. We applied this tool at both a species level and a clade level, and in both cases found that a significant number of gene families originally reported as accessory carried some level of spurious absences, often exceeding 50% of all reported events. To ensure that all accessory gene families used for downstream analysis were high-quality and biologically meaningful, we applied a conservative threshold of 80% agreement between the gene absences reported by Panaroo and those verified via BLAST, resulting in a total of 183 non-core, high-confidence gene families (**Figure 1C-D**). We expect that this conservative approach will generate false negatives, in the sense that some true accessory gene families will be excluded from our dataset. However, we also expect this approach to result in a lower false positive rate for the identification of accessory gene families, which is essential for our downstream analysis steps of ancestral reconstruction and inference of potential selective pressures.

Following these validation steps, accumulation curves of the pangenome at both the species and clade level showed diminishing returns, indicating that our dataset had likely sampled most of the gene families that exist in the isolates that have been sequenced for this species (**Figure 1E**), and demonstrates the highly clonal nature of each clade. In total, we identified 5693 gene families across our 695 isolates, with an average of 5620 families per genome (sd = 37.91). We found that 96.8% of gene families were categorized as core, appearing in at least 95% of isolates, with 78.2% of gene families appearing in 100% of isolates. We identified 127 validated accessory gene families that showed variation in at least one clade. 81 of these were unique to a single clade, with 27 were shared between exactly two clades, and 19 were shared between all three. However, we also identified 72 genes that were entirely clade-specific, appearing as core in some clades and absent in the remainder. Given that only about 21.8% of gene families exhibited

any variation in gene presence or absence, even when we designed our dataset to maximize genetic diversity, these results are consistent with a small effective population size of healthcare-associated isolates with little evidence of crossing. This is in contrast to yeast species where crossing is more common, such as *Aspergillus fumigatus,* where only 55-69% of the genome is core [26,29], or *Saccharomyces cerevisiae,* where ~60% of the genome is core [31].

The *C. auris* accessory genome shows repeated gene loss

The accessory genes present in this pangenome likely include genes and loci under selective pressure, which can cause the same gene family to be gained or lost multiple independent times. As our dataset draws heavily from clinical isolates, identification of the loci and genes with signatures of convergent evolution and selection in this dataset could provide insight into the selective pressures experienced by *C. auris* in healthcare settings. To accomplish this, we first generated a maximum-likelihood phylogeny tree with IQTREE [32], using only the single-copy core genes from the pangenome as input. We then subset our validated accessory genes based on the clade where they were identified (**Figure 2A-C**, top panels). We found that many gene families appeared on the same chromosome and exhibited strong linkage with each other, with identical patterns of presence and absence. To account for shared genomic rearrangements causing the loss of several genes at once, we clustered gene families together into single events if they occurred within 100 kb of each other and shared highly similar patterns of presence and absence (Jaccard index > 0.8, [33]). Of the 127 accessory genes, 51 could be grouped into clusters of two or more genes, with 17 clusters in total, resulting in 93 independent genetic loci. We then performed ancestral reconstruction on gene presence and absence data for all of the loci within each clade using phyloAMR [34] to identify the number of times each gene family was gained or lost (**Supplemental Figure 2A**). The number of transitions varied greatly between loci, even when both loci appeared in similar numbers of isolates. Depending on the clade, roughly 75-87% of loci had a high number of transitions via ancestral state reconstruction, with transitions equal to or greater than 80% of the number of isolates in which they appeared (**Supplemental Figure 2B**). In contrast, the remaining loci often exhibited significant clustering on the tree, with the gain or loss only appearing in specific subgroups (**Supplemental Figure 2A**). Contrary to our expectations, the loci with the highest numbers of transitions did not display any detectable enrichment near subtelomeric regions (**Figure 2A-C**, bottom panels), which are known to be more prone to genomic rearrangements and deletions [3,35,36]. However, the largest clusters of 4 or more linked gene families did all occur in subtelomeric regions, with the deletion of chromosome ends being responsible for the shared loss of several adjacent genes. Further examination of these large clusters of linked gene families revealed a large locus near the end of chromosome VI that was lost

multiple independent times in Clade IV (**Figure 2D**, **Supplementary Table 1**). Loss of this locus invariably caused truncations of the known adhesin gene, *ALS4112*, which is known to influence the clinically relevant trait of skin adhesion[7], as well as loss of two additional adhesins *IFF4109* and *IFF4110*[6]. Notably, this deletion only appeared in isolates derived from Rush University Medical Center, suggesting that this locus and phenotype may be under active selection in that environment.

We next extended our analysis to gene families that did not appear as accessory in any of the three clades, but still exhibited variation at the species level. This is possible when a gene appears in nearly every member of one or more clades and is entirely absent from the others. We found a total of 72 gene families that fell into this category. Clustering of linked genes reduced this number to 50 independent loci, with the largest cluster containing 5 linked gene families. Estimating the location of these loci revealed a nearly even distribution of genes across the genome (**Figure 2E**), similar to the clade-level accessory genes. The majority of these loci appeared in both Clade I and Clade III, while being absent in Clade IV, consistent with the known genetic distances between these clades [12].

Gene annotation data is limited in *C. auris*, but features such as InterProScan [37] and PFAM [38] terms can still provide insight into the putative functions and pathways enriched in sets of genes. We found significant enrichment for InterProScan terms related to hyphally-regulated cell wall proteins (IPR021031), major facilitator superfamily transporters including likely sugar transporters (IPR036259, IPR020846, IPR005828, IPR050360), PIK1 kinase (IPR021601, IPR049160), and RACK1 ribosomal subunit (IPR045223) among the accessory genes compared to the core genes (Bonferroni-corrected p-values < 0.05, only annotated genes tested). Significant PFAM terms across this same set of accessory genes were related to these functions, including PF11765 and PF15789 for hyphally-regulated cell wall proteins, PF00083 for sugar and other transporters, PF07690 for the major facilitator superfamily, and PF11522 and PF21245 for PIK1. Subsetting to accessory gene families within a clade did not have any significant impact on these terms.

While these analyses provide detailed insight into the genetic variation that exists in the accessory genome of *C. auris*, the highly clonal nature of our dataset means that we cannot assume that core gene families are conserved due to their biological functions. It is likely that the core gene families appearing in the 95% or more of our isolates consist of a combination of essential and nonessential genes, given how infrequently we observe presence and absence variation in our pangenome. This is supported by the fact that only 43.9% of the core genome from *C. auris* is present in at least two related species among *C. haemuli* and its nearest relatives (**Figure 3**). To further characterize and differentiate this large category of core gene families, we used the complementary approach of essentiality mapping via insertional mutagenesis. This will

allow us to better separate our large core gene families, as well as highlight unique *C. auris* biology and conserved druggable targets.

Transformation of a selection marker throughout the *C. auris* genome.

To perform insertional mutagenesis in *C. auris*, we developed and applied an insertional mutagenesis and mapping strategy. Precise genetic editing through transformation in *C. auris* is challenging, and many genetic cassettes end up incorporated to an off-target site instead of cleanly deleting a target gene [24,39]. To leverage off-target integration for mutagenesis, we used electroporation to incorporate a nourseothricin (NAT) resistance cassette throughout the genome, selected for NAT-resistant colonies in rich media, isolated genomic DNA, and performed library preparation for next-gen sequencing. We then adapted previous insertional mutagenesis techniques to identify insertion sites by digesting genomic DNA with the restriction enzyme MseI and ligating a U-linker to the newly exposed DNA end (**Supplemental Figure 3**). We applied this technique to the Clade I isolate AR0382, and found that the integration rate of this NAT cassette was high enough to produce insertional mutants without any additional external factors. We identified a total of 208,333 unique NAT cassette insertion sites across fifteen individual libraries, with an insertional density of approximately one insert per 59.5 bases. These insertion sites were widely distributed across all seven *C. auris* chromosomes (**Figure 4A**). We observed a strong bias towards the gene-deficient and repetitive region of rDNA repeats on the 3' end of chromosome 2 (CM076439.1). Previous insertional mutagenesis publications have noted these biases, which we accounted for by focusing our analysis on only genic NAT cassette inserts [15]. These NAT cassette integrations also displayed no detectable bias towards any sequence motif enrichment at the sequenced NAT 3' end. (**Figure 4B**). Whole genome sequencing of two individual strains also identified only a single insertion event per genome. Holistically, this analysis serves to validate the insert distribution of our method, showing that it is comparable to traditional transposon mutagenesis strategies. These results define a novel method for *C. auris* insertional mutagenesis that generates a pool of highly diverse insertional mutants for mapping and essentiality testing.

Essentiality mapping in *C. auris*

A central tenet of random mutagenesis essentiality mapping is that essential genetic regions will be insertion-depleted. As we increased our number of insertional libraries, we observed a saturation for the number of genes tolerating interruptions (**Figure 4D**). To determine if insertion sites were biased by chromatin accessibility, we optimized ATAC-seq for *C. auris* and integrated chromatin accessibility analysis on the AR0382 isolate, along with NAT insert site distribution. This allowed us to identify sites that were insert-poor, despite being within accessible regions. For

example, despite significant chromatin accessibility at gene loci *B9J08_04834* and *B9J08_04836*, we do not observe any insertions distributed across these genes. (**Figure 4C**). Additionally, these genes are flanked by predicted ammonium transmembrane transporter and ubiquitin protein ligase genes, which contain comparatively higher-frequency insertion sites in our sequencing. The orthologs to these genes have variable essentiality in other fungi, where *B9J08_04834* orthologs are broadly-essential and serve roles in mRNA 5' guanylyltransferase capping, and *B9J08_04836* is predicted to be part of the U2 snRNP complex, and is conserved essential in the *S. cerevisiae* Sigma1278b background [40].

When defining gene essentiality, we include several caveats. First, previous studies have found that some essential genes can tolerate interruptions, for example through alternative start sites or deletions past essential regions, so essential genes cannot solely be defined by zero-insert open reading frames [15,41]. Additionally, the lack of a defined insertion motif precludes us from using transposon mutagenesis analysis methods that require knowledge of insertion motifs to define insert-depleted regions [42]. Therefore, to conservatively predict gene essentiality, we developed a scoring system based on transgene insert frequency. We simulated repeated random redistribution of our NAT inserts (~115k) across the *C. auris* coding region. We then defined a z-score by normalizing the difference between our inserts for gene length, compared to the lower bound of the simulated insert distribution (Methods). To set a conservative cutoff for genes tolerating inserts, we aimed to define our group of putative essential genes as those that significantly deviated in insertion site count with a z-score cutoff of -5 or have zero inserts. We also provide all insert sites as a supplemental file **(Supplemental Data 1)**.

As an example of an essential gene, we can look at *C. auris* genes with widely-conserved essentiality in their orthologs, such as *TOR2* (**Figure 4E**). We observe significantly fewer insert sites in *TOR2* compared to the simulated distribution, although this number is not zero. In contrast, we observe that the number of observed inserts in other genes which we know are non-essential, such as *ALS4112* (*B9J08_04866*), does not deviate from a random distribution of inserts (**Figure 4F**). Although this analysis allows us to call essential *C. auris* genes without making other underlying assumptions, there are some limitations. For short genes, such as those < 700 bp, the lower bound of a simulated insert distribution includes zero inserts (**Figure 4G,H**), so it remains possible that we did not recover an insert due to chance. Therefore, we categorized these short genes without inserts as unresolvable (UN).

C. auris gene essentiality highlights divergent gene dependence.

Altogether, using these criteria, we established 614 genes as high confidence predicted essential and 219 as unresolvable (**Figure 5A, Supplemental Data 2**). We then compared these

predicted essential genes to the validated accessory genes from the pangenome. While essentiality can vary across different genetic backgrounds within the same species [40], we expected to see minimal overlap between the essential gene set and the validated accessory gene families. Consistent with our expectations, no validated accessory genes were categorized as essential genes in the insertional mutagenesis data. Additionally, we found that these essential genes were significantly enriched among the core genes that *C. auris* shares with *C. haemuli* and its close relatives, as opposed to the core genes that are not shared between these species (**Figure 3A**, Fisher's exact test pval = 3.3×10^{-5}).

For our total set of 614 putative essential genes, GO analysis highlighted enrichment for conserved essential functions (**Figure 5B**). Unresolved genes were similarly enriched for essential functions, including RNA polymerase activity, snRNA binding, and ribosome biogenesis (**Supplemental Figure 4A**). To address the lack of GO term annotation for *C. auris*, we also performed GO term enrichment using available *C. albicans* orthologs of our identified essential *C. auris* genes **(Supplemental Figure 4B).** This revealed additional enrichment in cytoskeleton organization, rRNA metabolism, transcription, translation, and DNA repair. Essential genes tend to exhibit higher expression [43,44]. Therefore, we compared the expression levels of our predicted essential and nonessential gene sets from a previously published RNAseq dataset[6] and observed that our predicted essentials have higher transcript levels (**Figure 5C**). Collectively, this supports our capacity to define essentiality and their correspondence with essential functions.

C. auris essentiality diverges from model yeasts

Since our permutation-based analysis does not incorporate information about essentiality from other species, we can then perform unbiased ortholog comparisons to identify evolutionary divergences in *C. auris* essentiality compared to related fungi. We compared essentiality and ortholog data to *S. cerevisiae* and *C. albicans,* where extensive work has been previously performed [40,45–48]. However, we recognize that some variability may be due to the conditions or strains under which essentiality was determined for each species. Overall, we found moderate agreement between these data sets, where 180/614 essential *C. auris* genes had essential orthologs in both *C. albicans* and *S. cerevisiae* (**Figure 5D**). We performed a chi-square test of independence to examine the conservation of essentiality across orthologs and observed a significant relationship ($\chi 2(1, N=614) = 114.5$, $p < 0.001$). However, we also found that a third of our putative essential genes in *C. auris* were not essential in *C. albicans* or *S. cerevisiae*. Though annotation of the divergently essential gene set was limited, GO term analysis revealed enrichment for DNA replication maintenance, oxidoreductase activity, and ion transport activity, suggesting enrichment for cell replication and cellular respiration (**Supplemental Figure 4C**). This is

consistent with the ability of both *C. albicans* and *S. cerevisiae* to grow anaerobically, while *C. auris* cannot [49]. Lastly, 47 of our putative *C. auris* essential genes do not have clear sequence orthologs to our two comparison species, precluding a comparative analysis of essentiality. All 47 of these lineage-restricted genes are core in Clade I, appearing in over 99% of isolates. We also observed transcripts (TPM >= 1) for each of these genes by RNAseq, confirming that these are actively expressed (**Supplemental Figure 4D**). These divergences in essentiality highlight the variable dependence on individual genes between these three fungal species and can provide a starting point for examining phenotypic variation between these species.

To experimentally test our inferences regarding one of these conserved essential *C. auris* genes, we chose one putative essential locus, the DASH complex subunit *Spc19* ortholog *B9J08_02579*, to put under the repressible tetOFF promoter [50]. The fungal-specific DASH complex associates with the kinetochore to ensure proper chromosome segregation during closed mitosis, though its essentiality depends on the timing and nature of kinetochore-microtubule interactions during the fungal cell cycle [51]. Therefore, if this gene is essential, we would expect to see cell cycle defects and culture stagnation, whereas if DASH is nonessential (as in *S. pombe*), loss of *spc19* would simply confer a decreased growth rate [52]. Initial growth on doxycycline-containing media displayed nearly complete growth inhibition, strongly supporting the inferred essentiality of this locus (**Figure 6A**), and consistency in essentiality between species. Through microscopic analysis of cell morphology, we observed that repression of Spc19 in *C. albicans* resulted in filaments consistent with cell cycle arrest (**Figure 6B**) [53,54]. Similarly, in *C. auris,* repression of Spc19 significantly inhibited growth and also resulted in elongated morphology, consistent with cell cycle arrest (**Figure 6B**) [55].

Our ortholog comparison analysis also revealed a set of 664 *C. auris* genes that were called nonessential under our scoring, despite having essential orthologs in related species. A number of the nonessential *C. auris* genes in this group can be explained by the difference in outgrowth between studies. In the studies conducted for gene essentiality in *C. albicans* and *S. cerevisiae*, mutants were grown in minimal media to identify genes for nutrient synthesis [41,46,56]. However, in our method, we outgrow using rich media, which allows for insertional mutagenesis of otherwise essential nutrient synthesis genes. To experimentally validate this, we generated a deletion strain of the *C. auris B9J08_01132* gene, a *LYS2* ortholog (**Figure 6C**). This strain exhibits auxotrophy on minimal media, which can be rescued through the supplementation of L-Lysine. This supports the sensitivity of our mutagenesis to properly identify gene nonessentiality in rich media, where conditionally essential genes like lysine synthesis are disposable.

However, we also identified cases where the discordance in gene essentiality suggests an evolutionary divergence in gene dependencies. To exemplify this, we generated a repressible

allele of the *SUI1 C. auris* ortholog, *B9J08_00976* [57]. In related species, Sui1 acts as translation initiation factor eIF1, which is an essential component for eukaryotic protein synthesis [58]. On doxycycline-containing media, we do not observe any growth defect in tetOFF-*00976*, contrasting with the lethal phenotype in an analogous tetOFF-*sui1 C. albicans* strain, despite strong repression through DOX treatment (**Figure 6D, Supplemental figure 5A**). The morphology of the tetOFF-*00976* was normal under doxycycline, but the tetOFF-*sui1 C. albicans* strain showed swelling and lack of growth **(Figure 6E).** The mechanism that allows for dispensable *SUI1* in *C. auris* is unclear, especially as the proteins show high sequence and structural similarity to the *C. albicans* protein **(Figure 6F)**, and we did not identify any paralogs or duplications through BLAST that would confer redundancy. Notably, a protein containing an SUI1-like domain was present in *C. auris,* but the rest of the protein structure was not shared **(Supplemental figure 5B).** Altogether, our experimental data supports a model where *C. auris* gene essentiality and translation biology has diverged from related species [56,59].

Lastly, identifying gene essentiality is key to future therapeutic development. To prioritize our *C. auris* gene essentiality set for for those that would make ideal drug targets, we performed a structural similarity search using FoldSeek to identify those lacking structural orthologs in human proteins (**Figure 6G**)[60], thus aiming to avoid targets with a high likelihood of concurrent host toxicity. From our search, we identified 6 *C. auris* genes that have high confidence structures (pLDDT >= 70), are predicted to be essential, and also lack a structural human ortholog (TM < 0.5). Of these six, three had already been identified as fungal-specific essential genes in this dataset, including *B9J08_04245*, a fungal mitochondrial subunit ortholog, and *B9J08_01210*, a tryptophan synthase ortholog. However, 209 of the *C. auris* essential proteins had low-confidence predicted structures, limiting our ability to look for structural orthologs.

Discussion:

In this study, we leverage both essentiality mapping and pangenome analysis to characterize the *Candidozyma auris* genome. Our pangenome analysis focuses on the identification of a high-confidence set of accessory gene families among a dataset of primarily clinical isolates, many of which exhibit strong phylogenetic signals and can be grouped into clusters of linked genetic loci. However, characterizing the *C. auris* genome through pangenomics is limited by the relatively small amount of genetic variation that exists in this highly clonal dataset, as well as sparse functional annotations for many of its genes. Additionally, our pangenome of mainly clinical and healthcare-adapted isolates may not sample all of the possible genetic variation possible in this species, especially given that the lack of environmental isolates suggests that there are unexplored reservoirs of *C. auris*. By taking advantage of the high off-target transformation rate

of *C. auris*, we developed and applied a novel insertional mutagenesis technique, mapping gene essentiality through a transposon-like method. In addition to identifying functions enriched among these essential genes, this addresses the limitations of the pangenome by dividing the core genes into essential and non-essential categories.

Inflation of pangenome size through the addition of spurious accessory or singleton gene families is a known problem in pangenomic studies [61]. These errors can be caused at multiple levels of the analysis, including problems caused by fragmented assemblies, technical artifacts of gene annotation, and issues with clustering genes into orthologous gene families. Given that our analysis focuses primarily on accessory genes, we chose to take a highly conservative approach to the accessory genome that minimized the number of false positives present in the data. This was done by only considering accessory genes for downstream analysis if two parallel methods, Panaroo and BLAST, consistently reported that the gene was absent from the same isolate genomes. While it is likely that this high level of stringency excludes some number of true accessory genes as false negatives, we found that this approach significantly improved the consistency of the pangenome with existing knowledge of *C. auris*. Validating our accessory genes in this way retained many of the gene loss events and excluded many of the gene gain events, which matches our expectations of a fungal species that does not demonstrate meiotic recombination [62]. This consistency also extends to our own essentiality mapping, where none of the accessory genes overlap with predicted essential genes.

We initially hypothesized that the highest-confidence accessory genes would be entirely localized to subtelomeric regions, as these locations are more likely to experience genomic instability and are the easiest to verify. However, we instead found that accessory genes are distributed roughly evenly across the genome, with no particular enrichment for any chromosomal regions. Additionally, we also found that over 63% of these accessory genes only varied within a single clade, being either absent or core in the other two. While the exact underlying genomic rearrangements causing these gene loss events are not yet known, this implies that our results cannot be explained simply by chromosomal instability in subtelomeric regions. This is further reinforced by our results from ancestral state reconstruction, where many accessory genes show very strong clustering and phylogenetic signals that are inconsistent with genomic instability. In future studies, we aim to combine this dataset with additional types of genetic variation, such as structural variants, SNPs, and indels, to better understand the genomic events leading to the presence and absence variation we observe here.

Pangenome analysis can also give insight into the evolutionary and ecological pressures that shape a species and the function of the core and accessory genomes. In pathogens, virulence genes in the accessory genome often allow for strain-specific pathogenicity mechanisms. As these

are more likely to be involved in host-pathogen arms races, they are generally under stronger positive selection and have distinct evolutionary trajectories compared to core conserved genes [63–65]. Additionally, core genes are more likely to be essential compared with accessory genes, although this essentiality can vary between strains, as the essentiality of a gene can be buffered by the presence of an accessory gene or other genetic background effect. This variability has been observed in other tractable systems; for example, gene essentiality between two strains of *S. cerevisiae,* S288c and Sigma1278b, showed only a 94% concordance [40]. Combining pangenome analysis with insertional mutagenesis in a set of *Streptococcus pneumoniae* strains revealed a set of 206 core and essential genes,186 core but not necessarily essential genes, and 128 genes that were accessory but still essential when present [66]. As our insertional mutagenesis strategy does not require prior genetic engineering of the *C. auris* strain, future work could include a strain or clade-level analysis of essentiality. Additionally, since our method of identifying *C. auris* essential genes is agnostic of orthology, we can identify differences in gene essentiality between orthologs in *S. cerevisiae* and the fungal pathogen *C. albicans,* and thus provide a platform for exploring species-specific biology.

In our set of essential proteins, we identified 6 proteins that had high structural quality scores but no obvious structural orthologs in humans. In this set were *C. auris* B9J08_04245, which shows homology to *S. cerevisiae* Nam9, a mitochondrial small ribosome subunit that mediates translation in the mitochondria through the mitoribosome. Previous work in *C. albicans* highlights the mitoribosome as a favorable therapeutic co-target, where deletion mutants confer fluconazole hypersensitivity [67]. Similarly, B9J08_01210 has homology to tryptophan synthase, an enzyme used by bacteria, plants, and fungi, but not mammals, for L-tryptophan synthesis, and has been considered a target for antimicrobial development [68–70]. However, it is important to note that 209 proteins had poor quality structure prediction, potentially due to a lack of training data for lineage-specific proteins in the AlphaFold models. As 32 of these proteins are also specific for *C. auris*, it is possible that these will also serve as good targets for future antifungal development, especially if they are essential in multiple clades of *C. auris*.

Although our novel method of non-homologous NAT cassette incorporation allows us to generate insertion mutant libraries with minimal reliance on exogenous genetic toolkits, there are other tradeoffs to our method. Specifically, unlike transposon-based mutagenesis that targets specific microsequences, such as *piggyBac* inserting at the TTAA motif, it is unclear whether the method of *C. auris* promiscuous insertion is sequence-dependent [71]. Since we do not have prior knowledge of the insertion space, we are restricted from performing typical transposon distribution analyses through Bayesian methods [72]. This requires a more conservative analysis of our insert distribution, because we cannot make as many assumptions about the underlying mechanism.

The promiscuity of NAT incorporation also starkly contrasts with the stability of the pangenome. In fact, promiscuous gene incorporation would suggest that gene gain events should be commonly observed. However, we see very little evidence of horizontal gene transfer or gene gain (singleton) events in the pangenome. We speculate that gene incorporation in *C. auris* may be a byproduct of genomic instability, such as high-frequency DNA breakage, that might be exacerbated by introduction of ssDNA carriers and high concentrations of linear NAT cassettes. Previous work in *C. auris* has hypothesized that promiscuous gene insertion may reduce on-target editing efficiency though DNA repair, such as non-homologous end joining [39]. However, knockouts of NHEJ proteins *KU70* and *LIG4* failed to improve editing efficiency, leaving it as an open question as to what is allowing for high rates of NAT incorporation. Considering that other mechanisms driving genetic diversity in *C. auris* are less clear, this instability potentially highlights a novel pathway for generating genetic variance [62]. A similar mechanism has been suggested in the fungal plant pathogens *Dothideomycetes*, where lack of the Holliday junction resolvase complex MUS81-MMS4 results in more frequent chromatid breakage [73]. While orthologs for MUS81-MMS4 are present in *C. auris*, further investigation of genome integrity pathways in *C. auris* may reveal the mechanism for its genomic plasticity.

Acknowledgements: We thank all members of the Snitkin and O'Meara labs for helpful comments. We thank MDHHS and Mary Hayden for sharing *C. auris* strains. Funding for this work was provided by:
AJL: MMMP T32AI007528, Michigan Postdoctoral Pioneer
TRO and ES: U19AI181767
TRO: R35GM147894
TRO: OORRS033123 University of Michigan
JRR: T32GM007544
GZ: F32AI181164, Michigan Postdoctoral Pioneer
LF: NIH R35GM128637, NIH R01AI134678

Methods:

Public Data and Sequencing

All public data used in this study was downloaded from the NCBI Pathogens database. Published SNP trees were used to select isolates from this database that differed by at least 30

SNPs. When possible, at least one isolate was selected from every SNP tree. All isolates not obtained from NCBI were sourced from either Rush University Medical Center or University of Michigan Medicine. Isolates from these sources were sequenced via Plasmidsaurus yeast hybrid sequencing orders, with library preparation performed as described below (section 'Library preparation and sequencing').

Assembly and Annotation

All genomes used in this study were assembled and annotated via the funQCD snakemake pipeline (https://github.com/Snitkin-Lab-Umich/funQCD). A detailed description of the steps and tools used is available on the GitHub repository. To briefly summarize, raw reads were cleaned via Trimmomatic and assembled via Spades. Assembled genomes were used as input for the fungal genome annotation tool Funannotate [26]. RNA-seq data from a panel of isolates was used as training data for structural annotation steps. GeneMark [74], InterProScan [37], and EggNOG-mapper [75] were used to improve functional annotations. These functional annotations are added coding sequences during the *de novo* assembly pipeline, and use protein homology to infer possible functions. Auriclass v0.5.4 (https://github.com/RIVM-bioinformatics/auriclass) was used to determine the clade of the isolate, and BUSCO [76] was used to evaluate completeness of both the annotation and assembly steps. MultiQC was used to aggregate QC data for downstream data cleaning.

Initial pangenome analysis

Pangenome analysis was performed with both Orthofinder [28] and Panaroo [27]. Protein .fasta files produced as output by Funannotate during the funQCD workflow were used as input for Orthofinder. Nucleotide .fasta and annotation .gff files from the same pipeline were used as input for Panaroo. As Panaroo was originally designed for bacterial genomes, it cannot process the ~5% of *C. auris* genes that contain introns. To account for this issue, we spliced introns out of these genes before providing input files to Panaroo. In the rare cases where introns could not be spliced due to overlapping another gene (<50 genes across all assemblies, <0.001%), the intron-containing gene was removed from analysis. Comparing Panaroo output before and after this splicing process verified that ~5% more genes per isolate were recovered, without significantly affecting the remaining gene families. Both Orthofinder and Panaroo were otherwise run with minimal changes from default settings.

BLAST Validation

All accessory and clade-specific gene families were validated via the pangenome_blast_smk pipeline (https://github.com/Snitkin-Lab-Umich/pangenome_blast_smk), primarily built around the command-line BLAST+ tool. A detailed description of the steps and functionality of this pipeline is available on GitHub. To briefly summarize, a custom BLAST database was first generated for all isolates. For each gene family, a single query sequence is extracted from an isolate that carries that gene, with priority given to high-quality hybrid assemblies. This query is BLASTed against the entire subject database, regardless of which isolates the gene originally appeared in the pangenome. Only hits with a sequence identity over 90%, query coverage over 90%, and e-values lower than 1e-5 are considered valid. In parallel, the pangenome graph from Panaroo is used to determine the closest neighbors of the target gene in each isolate, using synteny information. If within-clade accessory genes are being validated, BLAST hits are additionally required to appear on the same chromosome or contig within 50 kb of these expected neighboring genes. Species-wide searches do not use this threshold, as we observed less accurate synteny and neighbor information when performing cross-clade comparisons. BLAST hits are separated based on whether the gene family was originally reported as a presence or an absence in the pangenome for each isolate. Any gene families where more than 20% of the reported gene absences are found as valid hits via BLAST are considered spurious. As an additional requirement, at least one of the reported gene presences must be consistent with BLAST. Agreement between BLAST hits and pangenome presence data was used as a positive control, with roughly 2% of gene families displaying significant inconsistencies, indicating that the pipeline was performing accurately and as expected in this dataset.

Gene Family Location Estimation

Many gene families do not appear in reference genomes, even when using clade-specific references. For these gene families, the pangenome graph from Panaroo was used to infer its approximate genomic location. This was done using the find_pangenome_neighbors.py script within pangenome_blast_smk, as described above. To summarize, locations were determined by starting at the node representing the gene family of interest and traversing through the graph until a gene family was found that did appear in the reference genome. This process was repeated up to two times, depending on the number of edges connected to the node. If two neighbors were found, the approximate location of the target gene family was set as the interval between the two neighboring genes. If only one neighbor was found, the approximate location was set as the location of the only neighbor.

Strains and Culture Conditions

All strains used in this study are listed in **Supplemental Table S2**. Cells were cultured at 30°C in YPD or YNB base media, and liquid cultures were constantly agitated during incubation. When required, media was supplemented with 200 µg/mL nourseothricin, 40 µg/mL doxycycline, or 50 µg/mL L-Lysine. All strains were maintained as frozen stocks in 25% glycerol at 80°C. Microscopy was performed on a BioTek Lionheart FX Automated Microscope.

Primers and Plasmids

Supplemental tables contain lists of primers (**Supplemental Table S3**) and plasmids (**Supplemental Table S4**) used in this study. Plasmids constructed from fragments were assembled using the NeBuilder HIFI DNA Assembly Master Mix (NEB #E2621) according to the manufacturer manual.

***C. auris* Transformation**

C. auris transformation was performed as described previously [24]. Briefly, PCR was used to amplify linear cassettes of the repair, the targeting sgRNA, and Cas9. Each piece was purified using a Zymo DNA Clean & Concentrator kit (Cat no. D4034, Zymo Research) according to the manufacturer's instructions.

To prepare electro-competent cells, *C. auris* cells were grown overnight in YPD at 30 °C, pelleted, resuspended in TE buffer with 100 mM Lithium Acetate and incubated at 30 °C for 1 hr with constant shaking. Then, DTT was added at a final concentration of 25 mM and the cells were incubated at 30 °C for 30 min. Cells were harvested by centrifugation at 4 °C before being washed once with ice-cold water and once with ice-cold 1 M Sorbitol. Harvested cells were resuspended in ice-cold 1M sorbitol and kept on ice for transformation

For electroporation, 45 µL of competent cells were added to a pre-chilled 2 mm- gap electro-cuvette (Biorad) along with 500-1000 ng each of the linear cleaned DNA, and then cells were electroporated using a Bio-Rad MicroPulser Electroporator according to the pre-defined *P. pastoris* (PIC) protocol (2.0 kV, 1 pulse). The cells were then immediately recovered in 1 M cold Sorbitol, then moved to YPD and allowed to outgrow before selecting.

Strain Construction:

C. auris mutant strains were generated using a transient CRISPR system [24]. The universal Cas9 cassette was amplified from plasmid pTO135 using primers oTO143 and oTO41. Plasmids with repair cassettes were assembled using the NEBuilder HIFI DNA Assembly Master Mix (NEB #E2621) according to the manufacturer's instructions. Specific mutant colonies were identified through colony PCR using Phire Plant Direct PCR Master Mix (F160; Thermo Fisher Scientific)

according to the manufacturer's instructions. Each mutant was confirmed by at least three independent PCR reactions with distinct primer sets specific to the mutation site and compared to parental strains.

pTO505/tetO-B9J08_00976:

To generate t*etO-B9J08_00976* strains, the gRNA cassette was constructed by Splice-On-Extension PCR, incorporating a 20-bp guide sequence targeting the endogenous promoter (-203~-1) of *B9J08_00976* via primers oTO3255 and oTO3256. To generate plasmid with the repair template for t*etO-B9J08_00976*, the plasmid backbone was amplified from pTO139 using primers oTO590 and oTO591. The 500 bp upstream sequence (-703~-204) of *B9J08_00976* and the 500 bp gene sequence (+1~+500) were amplified from AR0382 genomic DNA using primer sets oTO3249/3250 and oTO3251/3252, respectively. The *tetO* cassette with a NAT selective marker was amplified from pTO240 with primers oTO3124 and oTO3125. The four fragments were assembled to generate a plasmid with the repair template for t*etO-B9J08_00976.* The repair cassette was amplified from pTO505 using primers oTO3259 and oTO3260 and transformed into AR0382 strain to generate t*etO-B9J08_00976* strains.

pTO487/tetO-B9J08_02579:

To generate t*etO-B9J08_02579* strains, the gRNA cassette was constructed by Splice-On-Extension PCR, incorporating a 20-bp guide sequence targeting the endogenous promoter (-223~-1) of *B9J08_02579* via primers oTO3074 and oTO3075. To generate plasmid with the repair template for t*etO-B9J08_02579*, the plasmid backbone was amplified from pTO139 using primers oTO590 and oTO591. The 500 bp upstream sequence (-735~-224) of *B9J08_02579* and the 500 bp gene sequence (+1~+501) were amplified from AR0382 genomic DNA using primer sets oTO3068/3069 and oTO3070/3071, respectively. The *tetO* cassette with a NAT selective marker was amplified from pTO240 with primers oTO3124 and oTO3125. The four fragments were assembled to generate a plasmid with the repair template for t*etO-B9J08_02579.* The repair cassette was amplified from pTO505 using primers oTO18 and oTO19 and transformed into AR0382 strain to generate t*etO-B9J08_02579* strains.

pTO540/tetO-B9J08_02579:

To generate *B9J08_01132::NAT* strains, the gRNA cassette was constructed by Splice-On-Extension PCR, incorporating a 20-bp guide sequence targeting the CDS of *B9J08_01132* via primers oTO3098 and oTO3099. To generate plasmid with the repair template for *B9J08_01132::NAT*, the plasmid backbone was amplified from pTO139 using primers oTO590 and

oTO591. The 500 bp upstream sequence (-1~-500) of *B9J08_01132* and the 500 bp gene sequence (+4177~+4677) were amplified from AR0382 genomic DNA using primer sets oTO3092/3093 and oTO3094/3095, respectively. The NAT selective marker was amplified from pTO128 with primers oTO668 and oTO1093. The four fragments were assembled to generate a plasmid with the repair template for *B9J08_01132::NAT.* The repair cassette was amplified from pTO540 using primers oTO18 and oTO19 and transformed into AR0382 strain to generate *B9J08_01132::NAT* strains.

Genomic DNA Isolation

Genomic DNA was isolated as previously described [6], using a PCA extraction method. Overnight cultures were harvested by centrifugation and resuspended in lysis buffer (2% (v/v) Triton X-100, 1% (w/v) SDS, 100 mM NaCl, 10 mM Tris-Cl, 1 mM EDTA). Cells were disrupted by bead-beating and DNA was extracted using PCA and chloroform, purified through ethanol precipitation and treated with RNase A (Qiagen, cat no. 19101). RNase was heat-inactivated, then extracted DNA was purified by ethanol precipitation and resuspended in water.

RNA extraction and RT-qPCR

RNA extraction and qRT-PCR was performed using a formamide extraction method, as previously described [6]. Cells were incubated with or without 40 ug/mL DOX overnight in YPD, harvested by centrifugation, and snap frozen at -80 °C before processing. For RNA extraction, pellets were resuspended in 100 µL FE Buffer (98% formamide, 0.01 M EDTA). 50 µL of 500 µm RNAse-free glass beads was added to this suspension and the mixture was homogenized for 30 sec 3 times using a BioSpec Mini-Beadbeater-16 (Biospec Products Inc., Bartlesville, OK, USA). The lysate was then clarified by centrifugation and RNA was purified using the Qiagen RNeasy mini kit (ref 74104, Qiagen) according to the manufacturer's instructions. Samples were DNAse treated with Qiagen DNAse (Qiagen, cat no. 79254).

ATACseq

To prepare spheroplasts, *C. auris* grown overnight in liquid culture was diluted and grown out for six hours to allow return to logarithmic growth. Approximately 1.6 OD of cells were harvested through centrifugation at 13,000 rpm for 1 minute. Supernatant was aspirated, and the pellet was washed in 1 mL ice cold water then centrifugation repeated at 13,000 rpm for one minute. Water was aspirated, pellet was resuspended in buffer 1 (100mM Tris pH 9.4, 10mM DTT) and incubated for 5 minutes at 30°C. Sample was then centrifuged at 2000x*g* for 5 minutes at 4°C,

and pellet was resuspended in cold buffer 2 (DPBS, 1 M sorbitol, 1.5 U zymolase [Zymo E100])). The sample was then incubated for 5 minutes at 30°C, harvested at 2000x*g* for 5 minutes at 4°C, and washed and centrifuged in 500 µL cold 1 M sorbitol without resuspending. Finally, spheroplasts were resuspended in 500 µL cold 1 M sorbitol, 50 µL was taken and harvested at 2000xg for 10 minutes at 4°C. Liquid was removed by pipette, and spheroplast was resuspended in tagmentation buffer according to the ActiveMotif ATAC-seq kit (#53150) method. After adding the tagmentation mix to the sample, we proceeded with the manufacturer's method starting with the Tagmentation Reaction and Purification section, to tagment samples and amplify sequencing libraries. Prepared ATACseq libraries were sequenced on either an Illumina Novaseq X or an Element AVITI24 at the University of Michigan Advanced Genomics Core. Fastq file quality was checked using FastQC and aligned to the *C. auris* B8441.v3 genome using the Burrows-Wheeler Aligner (BWA) MEM [77–79]. Bedgraph coverage files were created using Deeptools bamcoverage [80]. Peaks were identified using MACS3 [81].

Library preparation and sequencing

For insertional mutagenesis, *C. auris* AR0382 competent cells were transformed through electroporation as detailed above, using 600 ng of PCR amplified CaNAT from pTO128 using primers oTO668 and oTO1093. Transformed cells were allowed to outgrow in liquid YPD at 30°C for two hours, then nourseothricin stock was added to a final concentration of 200 µg/mL for an additional outgrowth up to 24 hours. Cells were then harvested through centrifugation for genomic DNA extraction.

To prepare libraries, we simultaneously prepared the genomic DNA and the custom U-linker, keeping components on ice. 1 µg of genomic DNA was digested with MseI (NEB #R0525) for three hours at 37°C in a benchtop thermocycler. Digested genomic DNA was then purified using a Zymo Research commercial clean and concentrate kit (Zymo #D4004) using the manufacturer specifications, with the added step of allowing the binding buffer to mix with the sample for three minutes at room temperature to inactivate the restriction enzyme. To prepare the U-linker (oTO2605), 10 µM oligo was incubated in an annealing buffer (100 mM NaCl, 10 mM Tris, pH 8, 1 mM EDTA, pH 8) for 90°C for two minutes, then cooled to 30°C at a ramp rate of -0.1°C/s. The reaction was then snap-cooled in ice and then additional reagents were added to simulate NEB 2.1 buffer (10 nM MgCl2, 100 µg/mL albumen). MseI was then added to the reaction, which was incubated at 37°C for two hours on a benchtop thermocycler. The reaction was purified using a Zymo oligo clean and concentrate kit, eluted with 21 µL of water, and 18 µL of U-linker was mixed with 2 µL 10x annealing buffer. This was heated on a thermocycler to 80°C for two minutes, then cooled to 23°C at a ramp rate of -0.1°C/s. The resulting digested U-linker reaction was kept on ice until ligation.

To ligate the digested U-linker and genomic DNA, we used the NEB quick ligation kit (NEB #M2200) to mix 2.5 µL of U-linker, 9 µL of digested genomic DNA, 12.5 µL of 2x quick ligation buffer, and 1 µL of ligase together, and incubated the reaction at room temperature for ten minutes. The reaction was stopped by adding 1 µL of 0.5 M EDTA. Ligated DNA was purified using Omega Bio-Tek Mag-Bind TotalPure SPRI beads (#M1378) at a 0.9x ratio of beads to reaction volume. We then PCR-amplified ligated DNA with a first round of PCR to add Illumina sequences with the following master mix: 12.5 µL milli-Q water, 10 µL 5x Q5 reaction buffer, 2.5 µL of 10 µM oTO2516, 2.5µL of 10 µM oTO2517, 1 µL of 25mM dNTP mix, 0.5 µL Hot-start Q5 Polymerase, and 1 µL USER enzyme. 30 µL of master mix was added to 20 µL of ligated DNA, for the following reaction: initial incubation at 37°C for USER activity, an initial denaturation at 95°C for 1 minute, then 12 cycles of denaturation (95°C, 15 seconds), annealing (63°C, 15 seconds), and extension (72°C, 1 minute); this was followed by a final extension cycle (72°C, 2 minutes). The product of this reaction was purified using Omega Bio-Tek Mag-Bind TotalPure SPRI beads (#M1378) at a 0.9x ratio of beads to reaction volume. Purified DNA was then added to a reaction master mix (10 µL 5x Q5 reaction buffer, 1 µL of 25mM dNTP mix, 8.5 µL milli-Q water, 0.5 µL Hot-start Q5 Polymerase, 5µL of 10 µM labeled i5 sequencing primer, 5µL of 10 µM labeled i7 sequencing primer) and run on a Bio-Rad T100 Thermocycler for the following reaction: initial incubation at 37°C for USER activity, an initial denaturation at 98°C for 1 minute, then 5 cycles of denaturation (98°C, 10 seconds), annealing (65°C, 40 seconds), and extension (72°C, 40 seconds); this was followed by a final extension cycle (72°C, 3 minutes). This final library was purified using Omega Bio-Tek Mag-Bind TotalPure SPRI beads (#M1378) at a 0.9x ratio of beads to reaction volume. Libraries were sequenced on either an Illumina NovaSeqX or Element AVITI24 platform through the University of Michigan Advanced Genomics Core.

Library analysis

Library sequencing quality control was performed by the University of Michigan Advanced Genomics Core. Read quality was confirmed using FastQC and sequences containing the end of the NAT cassette were identified by initially aligning to pTO128 using the Burrows-Wheeler Aligner (BWA) MEM with minimum seed length (-k) 19 and a band width (-w) of 2. Soft clipped reads corresponding to the junction between the NAT cassette and *C. auris* genomic DNA were then extracted using extractSoftClipped from SE-MEI (Dpryan, https://github.com/dpryan79/SE-MEI). Extracted sequences were then aligned to *C. auris* GCA_002759435.3_Cand_auris_B8441_V3 assembly using BWA MEM to identify NAT insertion site regions [77,78]. Insert sites were processed using bedtools to condense regions to single-nucleotide resolution at the junction site, and adjacent insertion sites were merged. Insert motif enrichment analysis was performed using the MEME suite of tools on extracted genomic sequences fifteen bases upstream and downstream of NAT insertion

sites [82]. Analysis of NAT insert regions was performed in R and visualized using ggplot2 (4.0.0) and karyoploteR [83]. For the permutation analysis, shorter genes were categorized as unresolved if their permutation lower bound was zero. Some genes were masked based on their proximity to highly repetitive telomeric regions, and include B9J08_03013-03019 from the repetitive telomeric region of CM076439.1, and B9J08_04428 on the telomeric region of CM076441.1. 5586 remaining total annotated genes were considered for this analysis. Gene Ontology enrichment was performed through the Candida Genome Database [84].

Gene Expression analysis

Previously published RNAseq data from Santanta et al. was used in this study [6]. Raw reads were assessed for quality using FastQC and trimmed to remove poor quality reads or adapter content using Trimmomatic [79,85]. Trimmed reads were mapped to the *C. auris* B8441 reference transcriptome (NCBI GCA_002759435.3) and quantified using salmon v1.9 [86]. Read counts normalized to transcripts per million (TPM) were used for the subsequent analysis.

FIGURE LEGENDS

Figure 1: Refining the *C. auris* pangenome. a, Metrics of the *de novo* assemblies used as input for pangenome analysis, following all QC steps. Top left: Nucleotide BUSCO score. Bottom left: Protein BUSCO score. Top right: N50. Bottom right: Number of contigs. **b,** Alluvial plot displaying the difference in gene family category between pangenomes produced by Orthofinder and Panaroo, when both tools are run on the same dataset of 694 *C. auris* assemblies. “Removed” indicates the corresponding gene family was not found in the pangenome. **c,** Number of core and non-core gene families present in the pangenome after BLAST validation was performed. **d,** *C. auris* pangenome after all validation steps were performed. Only non-core gene families are shown. Gene families are ordered by decreasing frequency from right to left. Isolates are ordered by clade on the y-axis. **e,** Accumulation curves for the post-validation pangenome. Clade-specific curves are shown in addition to the species-wide curve.

Figure 2: Characterization of the *C. auris* accessory genome. a, Top panel, left: Maximum-likelihood phylogeny tree for Clade I isolates. Top panel, right: Presence/absence matrix for all gene families that are accessory in Clade I, appearing in less than 95% of isolates but more than 1. Gene families are ordered on the x-axis by estimated position, based on synteny information from the pangenome. Bottom panel: Genomic location, number of isolates, and number of transitions for the

accessory genes shown on the top panel. Number of transitions were determined via ancestral state reconstruction of the gene presence/absence matrix. Gene families exhibiting strong linkage are clustered together and treated as a single locus. **b,** Similar to panel A, with isolates and gene families from Clade III. **c,** Similar to panel A, with isolates and gene families from Clade IV. **d,** Representative isolates showing two different truncations of the *ALS4112* adhesin gene. All isolates shown were drawn from the Rush University Medical Center dataset. **e,** Estimated locations of clade-specific genes. These gene families are core in one or more clades and absent from the remainder.

Figure 3: *C. auris* core genome compared to *C. haemuli*. a, Results of a pangenome generated using 703 genomes from all six clades of *C. auris*, as well as reference genomes for *Candida haemuli*, *Candida duohaemuli*, and *Candida pseudohaemuli*. For each clade, 2-3 representative isolates are shown, prioritizing the highest-quality assemblies when possible. Only gene families categorized as core (present in >95% of isolates) are shown on the heatmap, and are ordered on the x-axis based on decreasing frequency from left to right.

Figure 4: Transformation of a selection marker throughout the *C. auris* genome. a, Karyoplot illustrating NAT-resistance gene insertion density (red) per 50kb-bin across the *C. auris* genome at over 200k unique insertion sites. Gene density is shown below (gray). 15kb on chromosome CM076439.1 was masked to improve visible scale. **b,** Sequence logo generated from the hybrid NAT-gDNA insertions. 'CATCCAGT' corresponds to the end of the NAT resistance cassette. **c,** IGV tracks displaying select *C. auris* genes at the indicated coordinates (top). Insert sites, ATAC-seq, and MACS3-identified ATAC-seq peaks are shown below. **d,** Number of gene ORFs containing at least one NAT insert site reached an asymptote with additional libraries. **(e,f,g,h,)** (top) Random permutation simulation analysis distributed the proportion of inserts landing in coding ORFs (~115k sites) across the gene sample space. Comparing the lower bound of the distribution to our observation (red line) allowed us to further identify ORFs of essential genes. Inserts were normalized by gene length. (below) IGV tracks displaying select C. auris genes, insert sites, chromatin accessibility (ATACseq), and transcription (RNAseq) at the indicated coordinates.

Figure 5: *C. auris* gene essentiality highlights divergent gene dependence. a, *C. auris* genes that are essential, nonessential, or unresolved by permutation analysis. **b,** GO functional analysis of 614 essential *C. auris* genes, performed using the Candida Genome Database GO Term Finder. **c,** RNAseq expression in log2(TPM+1) for *C. auris* essential and nonessential genes. Unresolved genes were binned with nonessential genes. Data from Santana et al. 2023. **d,** Comparison of 614 *C. auris* essential genes to orthologs in *C. albicans* and *S. cerevisiae*. 47 *C. auris* putative essential

genes had no annotated ortholog to either related species, and are a subset of the previous group of 252 essential *C. auris* genes that are nonessential or have no ortholog in *C. albicans* and *S. cerevisiae*.

Figure 6: Validation of *C. auris* gene essentiality. a, IGV tracks for the *SPC19* locus and a spot plate showing growth for the indicated strains under a tetOFF promoter in the presence or absence of doxycycline. **b,** Morphology of the tetOFF-SPC19 strains *C. albicans* and *C. auris* in the presence or absence of doxycycline in YPD. **c,** IGV tracks at the *LYS2* locus and a YNB agar spot plate showing growth in the presence or absence of lysine. **d,** IGV tracks for the *SUI1* locus and a spot plate showing growth for the indicated strains under a tetOFF promoter in the presence or absence of doxycycline. **e,** Morphology of the indicated strains in the presence or absence of doxycycline. **f,** Alphafold-predicted structures of the *C. auris* or *C. albicans* Sui1 proteins. **g,** Structural analysis of *C. auris* essential genes identifies candidates without structural orthologs in humans.

SUPPLEMENTARY FIGURE LEGENDS AND TABLES

Supplementary Figure 1. Pangenome graph. **a,** Representative example of the pangenome graph produced by Panaroo. Each node represents a single gene family, regardless of how many isolates it appears in. Edges connect gene families if they are adjacent to each other in at least one assembly. This information was used to determine the gene adjacency information for the BLAST validation pipeline.

Supplementary Figure 2. Ancestral state reconstruction. **a,** Representative example of a validated accessory gene from Clade I that shows a strong phylogenetic signal. Most gene loss events cluster tightly on the tree. A total of 55 isolate genomes lack this gene family. **b,** Representative example of a validated accessory gene from Clade I that shows a near-random distribution of gene loss events when compared to the phylogenetic tree. Despite the difference in distribution compared to panel A, nearly the same number of isolates lack this gene family (58).

Supplementary Figure 3. Design for unguided integration mutagenesis using a NAT selection marker. The NAT resistance cassette was transformed into competent *C. auris* cells via electroporation. Genomic DNA was harvested, digested with a MseI restriction enzyme, and ligated to a U-linker to make linear hybrid junction fragments. These fragments were then amplified through two steps of PCR to produce Illumina-compatible sequencing libraries.

Supplementary Figure 4. (a,b,c), GO functional analysis of **a,** 219 *C. auris* genes in our defined unresolved category, **b,** 511 *C. auris* essential genes with orthologs in *C. albicans* and **c,** 253 *C. auris* essential genes without essential orthologs in *C. albicans* and *S. cerevisiae*. GO enrichment analysis was performed using the Candida Genome Database GO Term Finder. **d,** RNAseq expression for *C. auris* essential and nonessential genes. Unresolved genes were binned with nonessential genes. Set of 47 *C. auris* essential genes with no orthologs in *C. albicans* and *S. cerevisiae* are labeled in blue. Data from Santana et al. 2023.

Supplementary Figure 5. (a), qRT-PCR for *B9J08_00976* in indicated *C. auris* isolates, with and without 40 µg/mL doxycycline treatment. **(b)** Alphafold predicted structures for the *C. auris* and *C. albicans* Sui1 protein in comparison with B9J08_00176, which has a Sui1-like fold. RMSD over the shared domain is 0.838 Å.

Supplementary tables

Table S1	ALS4112 deletions among Clade IV isolates from Rush University Medical Center.
Table S2	Strains used in this study
Table S3	Oligonucleotides used in this study
Table S4	Plasmids used in this study

References:

1. Santana, D. J., Zhao, G. & O'Meara, T. R. The many faces of Candida auris: Phenotypic and strain variation in an emerging pathogen. *PLoS Pathog.* **20**, e1012011 (2024).
2. Jacobs, S. E. *et al.* Candida auris pan-drug-resistant to four classes of antifungal agents. *Antimicrob. Agents Chemother.* **66**, e0005322 (2022).
3. Muñoz, J. F. *et al.* Clade-specific chromosomal rearrangements and loss of subtelomeric adhesins in Candida auris. *Genetics* (2021) doi:10.1093/genetics/iyab029.
4. Bravo Ruiz, G. *et al.* Rapid and extensive karyotype diversification in haploid clinical Candida auris isolates. *Curr. Genet.* **65**, 1217–1228 (2019).

5. Burrack, L. S., Todd, R. T., Soisangwan, N., Wiederhold, N. P. & Selmecki, A. Genomic diversity across Candida auris clinical isolates shapes rapid development of antifungal resistance in vitro and in vivo. *MBio* **13**, e0084222 (2022).
6. Santana, D. J. *et al.* A Candida auris-specific adhesin, Scf1, governs surface association, colonization, and virulence. *Science* **381**, 1461–1467 (2023).
7. Zhao, G. *et al.* Adhesin Als4112 promotes Candida auris skin colonization through interactions with keratinocytes and extracellular matrix proteins. *Nat. Commun.* **16**, 5673 (2025).
8. Brandt, P. *et al.* High-throughput profiling of Candida auris isolates reveals clade-specific metabolic differences. *Microbiol. Spectr.* **11**, e0049823 (2023).
9. Khan, Z. *et al.* Increasing prevalence, molecular characterization and antifungal drug susceptibility of serial Candida auris isolates in Kuwait. *PLoS One* **13**, e0195743 (2018).
10. Vallabhaneni, S. *et al.* Investigation of the First Seven Reported Cases of Candida auris, a Globally Emerging Invasive, Multidrug-Resistant Fungus - United States, May 2013-August 2016. *MMWR Morb. Mortal. Wkly. Rep.* **65**, 1234–1237 (2016).
11. Proctor, D. M. *et al.* Integrated genomic, epidemiologic investigation of Candida auris skin colonization in a skilled nursing facility. *Nat. Med.* (2021) doi:10.1038/s41591-021-01383-w.
12. Chow, N. A. *et al.* Tracing the evolutionary history and global expansion of Candida auris using population genomic analyses. *MBio* **11**, e03364–19 (2020).
13. Lionakis, M. S. & Chowdhary, A. *candida auris* infections. *N. Engl. J. Med.* **391**, 1924–1935 (2024).
14. Gale, A. N. *et al.* Identification of Essential Genes and Fluconazole Susceptibility Genes in Candida glabrata by Profiling Hermes Transposon Insertions. *G3* **10**, 3859–3870 (2020).
15. Billmyre, R. B. *et al.* Landscape of essential growth and fluconazole-resistance genes in the human fungal pathogen Cryptococcus neoformans. *PLoS Biol.* **23**, e3003184 (2025).
16. Segal, E. S. *et al.* Gene Essentiality Analyzed by In Vivo Transposon Mutagenesis and Machine Learning in a Stable Haploid Isolate of Candida albicans. *MBio* **9**, (2018).
17. Xu, T. *et al.* The identification of essential cellular genes is critical for validating drug targets. *Drug Discov. Today* **29**, 104215 (2024).
18. Poulsen, B. E. *et al.* Defining the core essential genome of Pseudomonas aeruginosa. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 10072–10080 (2019).
19. Torres, M., Paszti, S. & Eberl, L. Shedding light on bacteria-host interactions with the aid of TnSeq approaches. *MBio* **15**, e0039024 (2024).
20. Freed, N. E., Bumann, D. & Silander, O. K. Combining Shigella Tn-seq data with gold-standard E. coli gene deletion data suggests rare transitions between essential and non-essential gene functionality. *BMC Microbiol.* **16**, 203 (2016).

21. Fouts, D. E. *et al.* Integrating genomic and Tn-Seq data to identify common in vivo fitness mechanisms across multiple bacterial species. *MBio* **16**, e0198825 (2025).
22. Gao, J. *et al.* LncRNA DINOR is a virulence factor and global regulator of stress responses in Candida auris. *Nat. Microbiol.* **6**, 842–851 (2021).
23. Song, Y. *et al.* Candida auris vacuolar calcium pump mediates fluconazole efflux and resistance evolution. *Nat. Microbiol.* **11**, 786–801 (2026).
24. Santana, D. J. & O'Meara, T. R. Forward and reverse genetic dissection of morphogenesis identifies filament-competent Candida auris strains. *Nat. Commun.* **12**, 7197 (2021).
25. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Res.* **10**, 33 (2021).
26. Lofgren, L. A., Ross, B. S., Cramer, R. A. & Stajich, J. E. The pan-genome of Aspergillus fumigatus provides a high-resolution view of its population structure revealing high levels of lineage-specific diversity driven by recombination. *PLoS Biol.* **20**, e3001890 (2022).
27. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).
28. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, (2019).
29. Barber, A. E. *et al.* Aspergillus fumigatus pan-genome analysis identifies genetic variants associated with human infection. *Nat Microbiol* **6**, 1526–1536 (2021).
30. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
31. Loegler, V. *et al.* From genotype to phenotype with 1,086 near telomere-to-telomere yeast genomes. *Nature* 1–10 (2025).
32. Minh, B. Q. *et al.* IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
33. Jaccard, P. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.[1]. *New Phytol.* **11**, 37–50 (1912).
34. Gontjes, K. J. *et al.* Phylogenetic context of antibiotic resistance provides insights into the dynamics of resistance emergence and spread. *medRxiv* (2025) doi:10.1101/2025.06.04.25328982.
35. Castaño, I. *et al.* Telomere length control and transcriptional regulation of subtelomeric adhesins in Candida glabrata. *Mol. Microbiol.* **55**, 1246–1258 (2005).
36. Chow, E. W. L., Morrow, C. A., Djordjevic, J. T., Wood, I. A. & Fraser, J. A. Microevolution of Cryptococcus neoformans driven by massive tandem gene amplification. *Mol. Biol. Evol.* **29**, 1987–2000 (2012).
37. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics*

30, 1236–1240 (2014).

38. Paysan-Lafosse, T. *et al.* The Pfam protein families database: embracing AI/ML. *Nucleic Acids Res.* **53**, D523–D534 (2025).
39. Sofras, D. *et al.* A systematic comparison of CRISPR-Cas9 allele editing in Candida auris demonstrates unreliable cassette integration and effective episomal plasmid-based editing. *Sci. Rep.* **15**, 35105 (2025).
40. Dowell, R. D. *et al.* Genotype to phenotype: a complex problem. *Science* **328**, 469 (2010).
41. Chen, P., Michel, A. H. & Zhang, J. Transposon insertional mutagenesis of diverse yeast strains suggests coordinated gene essentiality polymorphisms. *Nat. Commun.* **13**, 1490 (2022).
42. DeJesus, M. A. *et al.* Bayesian analysis of gene essentiality based on sequencing of transposon insertion libraries. *Bioinformatics* **29**, 695–703 (2013).
43. Lloyd, J. P., Seddon, A. E., Moghe, G. D., Simenc, M. C. & Shiu, S.-H. Characteristics of plant essential genes allow for within- and between-species prediction of lethal mutant phenotypes. *Plant Cell* **27**, 2133–2147 (2015).
44. Liang, Y.-T., Luo, H., Lin, Y. & Gao, F. Recent advances in the characterization of essential genes and development of a database of essential genes. *Imeta* **3**, e157 (2024).
45. Fu, C. *et al.* Expansion of the functional genomics GRACE library reveals genes relevant for temperature-dependent fitness in Candida albicans. *PLoS Biol.* **23**, e3003409 (2025).
46. O'Meara, T. R. *et al.* Global analysis of fungal morphology exposes mechanisms of host cell escape. *Nat. Commun.* **6**, 6741 (2015).
47. Roemer, T. *et al.* Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery. *Mol. Microbiol.* **50**, 167–181 (2003).
48. Cherry, J. M. *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40**, D700–5 (2012).
49. Day, A. M., McNiff, M. M., da Silva Dantas, A., Gow, N. A. R. & Quinn, J. Hog1 regulates stress tolerance and virulence in the emerging fungal pathogen Candida auris. *mSphere* **3**, e00506–18 (2018).
50. Iyer, K. R., Kim, S. H., Robbins, N. & Cowen, L. E. Downregulation of essential genes in the fungal pathogen Candida auris. *Methods Mol. Biol.* **2517**, 111–126 (2022).
51. Thakur, J. & Sanyal, K. The essentiality of the fungus-specific Dam1 complex is correlated with a one-kinetochore-one-microtubule interaction present throughout the cell cycle, independent of the nature of a centromere. *Eukaryot. Cell* **10**, 1295–1305 (2011).
52. Rodríguez-López, M. *et al.* Broad functional profiling of fission yeast proteins using phenomics and machine learning. *Elife* **12**, (2023).

53. Senn, H., Shapiro, R. S. & Cowen, L. E. Cdc28 provides a molecular link between Hsp90, morphogenesis, and cell cycle progression in Candida albicans. *Mol. Biol. Cell* **23**, 268–283 (2012).

54. Shapiro, R. S., Robbins, N. & Cowen, L. E. Regulatory circuitry governing fungal development, drug resistance, and disease. *Microbiol. Mol. Biol. Rev.* **75**, 213–267 (2011).

55. Kim, S. H. *et al.* Genetic Analysis of Candida auris Implicates Hsp90 in Morphogenesis and Azole Tolerance and Cdr1 in Azole Resistance. *MBio* **10**, e02529–18 (2019).

56. Fu, C. *et al.* Leveraging machine learning essentiality predictions and chemogenomic interactions to identify antifungal targets. *Nat. Commun.* **12**, 6497 (2021).

57. Park, Y.-N. & Morschhäuser, J. Tetracycline-inducible gene expression and gene deletion in Candida albicans. *Eukaryot. Cell* **4**, 1328–1342 (2005).

58. Naranda, T., MacMillan, S. E., Donahue, T. F. & Hershey, J. W. SUI1/p16 is required for the activity of eukaryotic translation initiation factor 3 in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **16**, 2307–2313 (1996).

59. Iyer, K. R. *et al.* Translation inhibition by rocaglates activates a species-specific cell death program in the emerging fungal pathogen Candida auris. *MBio* **11**, (2020).

60. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).

61. Tonkin-Hill, G., Corander, J. & Parkhill, J. Challenges in prokaryote pangenomics. *Microb. Genom.* **9**, (2023).

62. Ross, Z. K. & Lorenz, A. Is Candida auris sexual? *PLoS Pathog.* **16**, e1009094 (2020).

63. Plissonneau, C., Hartmann, F. E. & Croll, D. Pangenome analyses of the wheat pathogen Zymoseptoria tritici reveal the structural basis of a highly plastic eukaryotic genome. *BMC Biol.* **16**, 5 (2018).

64. Croucher, N. J. *et al.* Diverse evolutionary patterns of pneumococcal antigens identified by pangenome-wide immunological screening. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E357–E366 (2017).

65. Allen, J. P., Snitkin, E., Pincus, N. B. & Hauser, A. R. Forest and Trees: Exploring Bacterial Virulence with Genome-wide Association Studies and Machine Learning. *Trends Microbiol.* **29**, 621–633 (2021).

66. Rosconi, F. *et al.* A bacterial pan-genome makes gene essentiality strain-dependent and evolvable. *Nat. Microbiol.* **7**, 1580–1592 (2022).

67. Sun, N., Parrish, R. S., Calderone, R. A. & Fonzi, W. A. Unique, diverged, and conserved mitochondrial functions influencing Candida albicans respiration. *MBio* **10**, (2019).

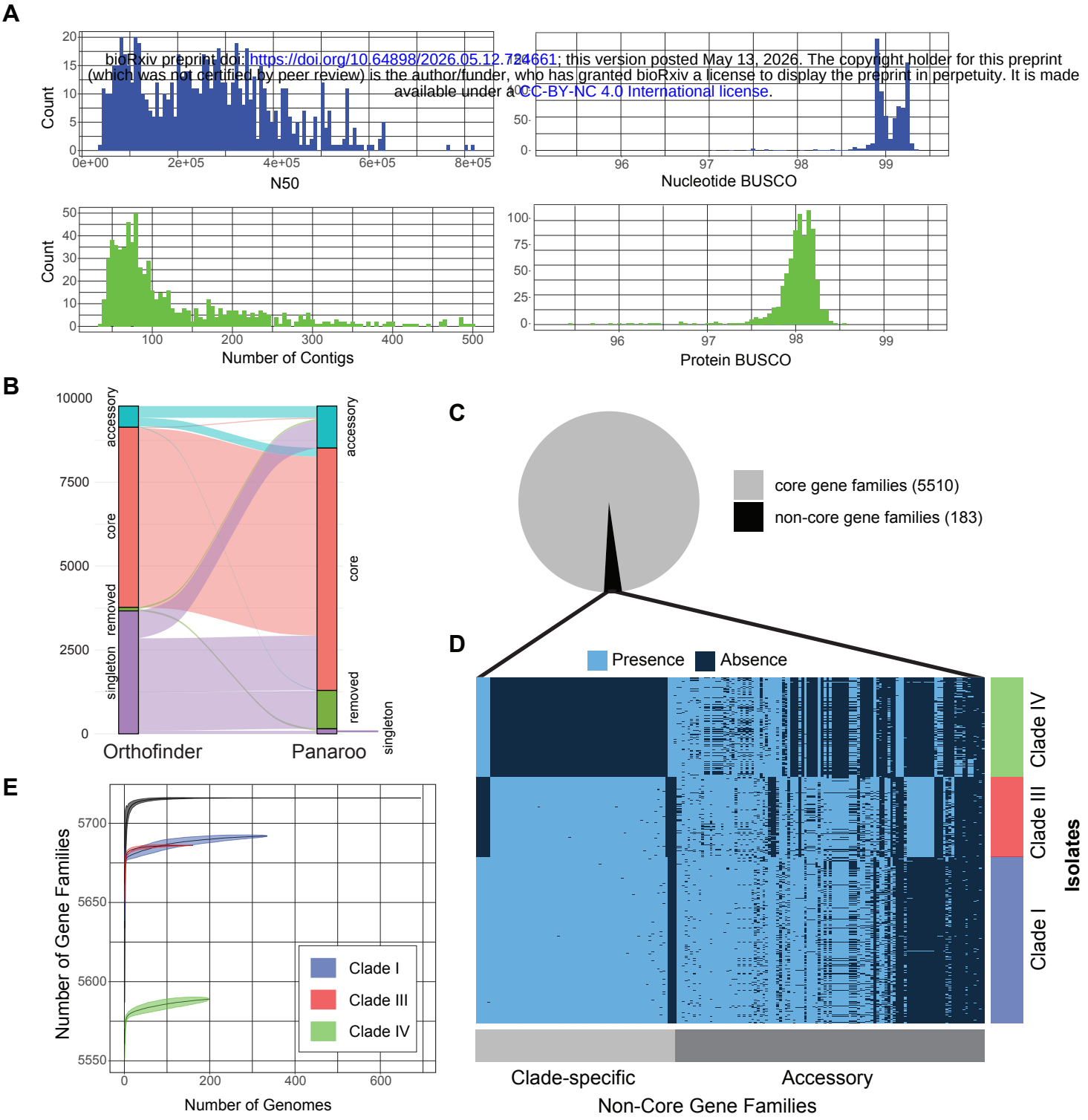
68. Bosken, Y. K. *et al.* Discovery of antimicrobial agent targeting tryptophan synthase. *Protein*

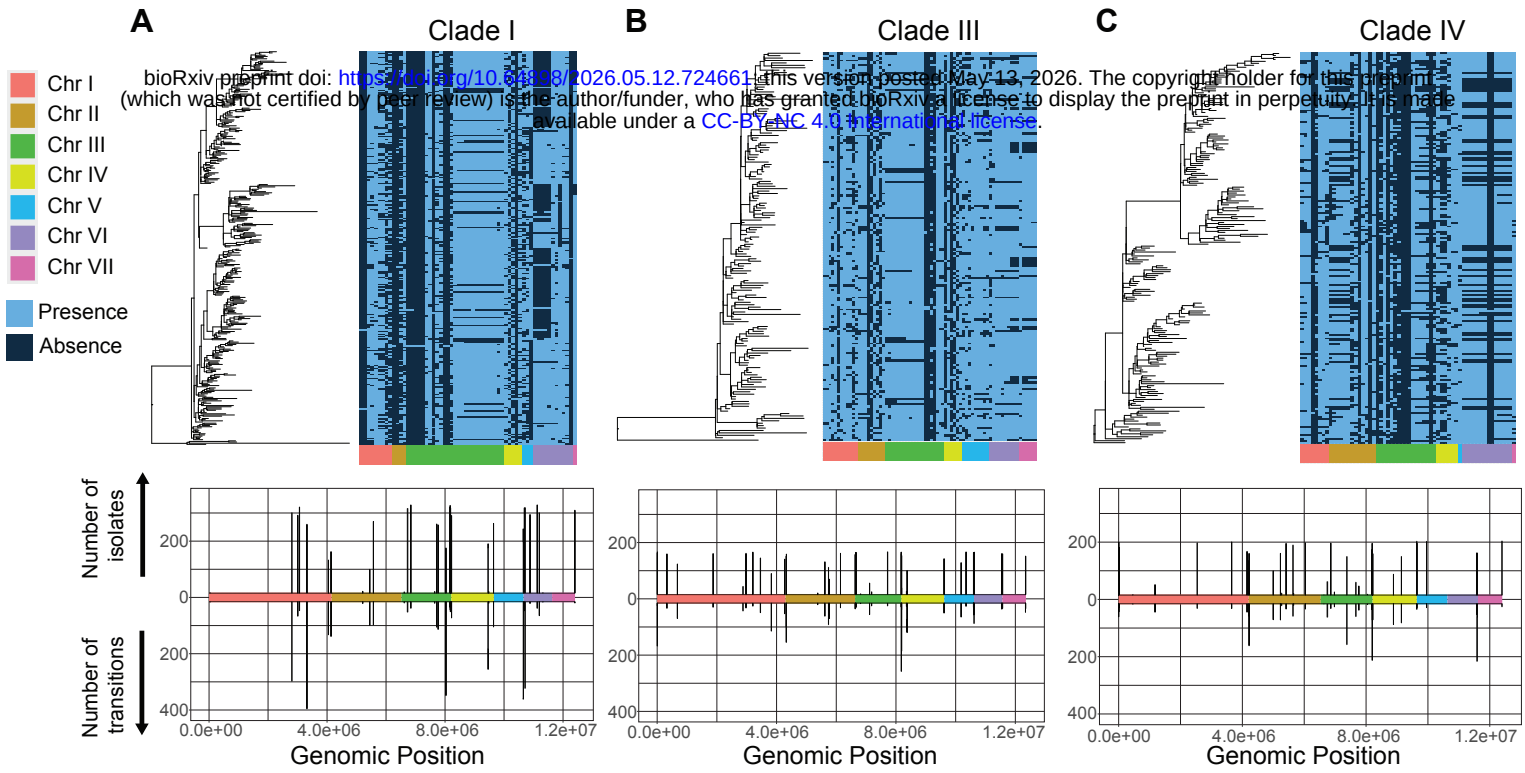
Sci. **31**, 432–442 (2022).

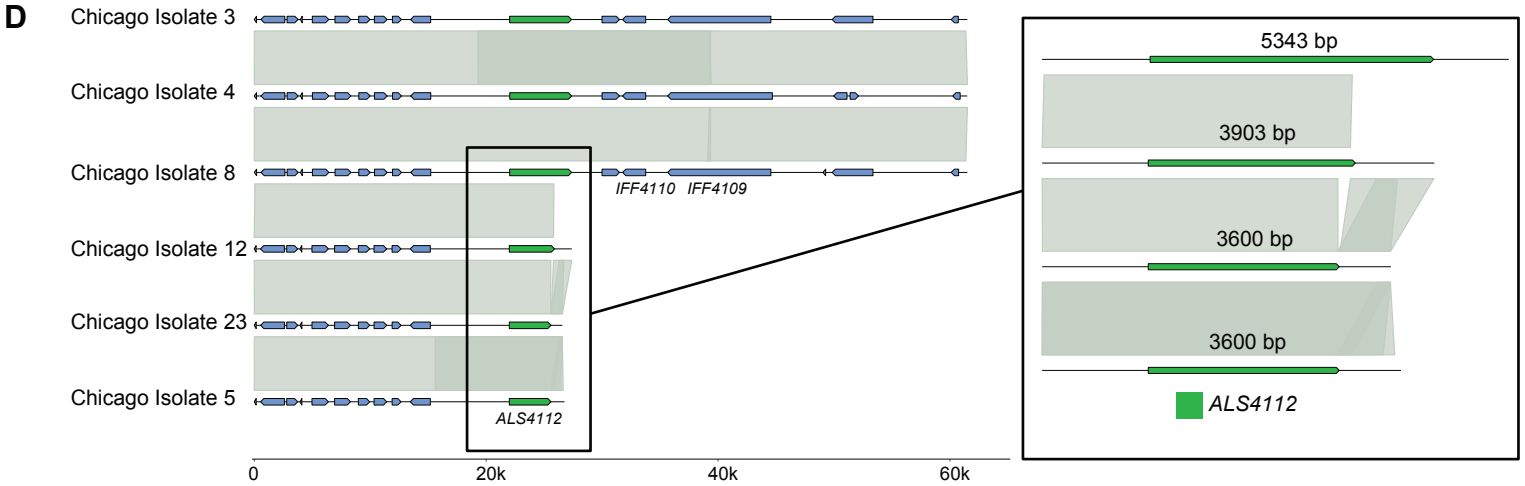
69. Abrahams, K. A. *et al.* Inhibiting mycobacterial tryptophan synthase by targeting the inter-subunit interface. *Sci. Rep.* **7**, 9430 (2017).
70. Choera, T., Zelante, T., Romani, L. & Keller, N. P. A multifaceted role of tryptophan metabolism and indoleamine 2,3-dioxygenase activity in Aspergillus fumigatus-host interactions. *Front. Immunol.* **8**, 1996 (2017).
71. Zhao, S. *et al.* PiggyBac transposon vectors: the tools of the human gene encoding. *Transl. Lung Cancer Res.* **5**, 120–125 (2016).
72. DeJesus, M. A., Ambadipudi, C., Baker, R., Sassetti, C. & Ioerger, T. R. TRANSIT--A software tool for Himar1 TnSeq analysis. *PLoS Comput. Biol.* **11**, e1004401 (2015).
73. Covo, S. Genomic instability in fungal plant pathogens. *Genes (Basel)* **11**, 421 (2020).
74. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506 (2005).
75. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. EggNOG-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
76. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
77. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
78. Cauldron, N. C., Shea, T. & Cuomo, C. A. Improved genome assembly of Candida auris strain B8441 and annotation of B11205. *Microbiol. Resour. Announc.* **13**, e0051224 (2024).
79. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
80. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
81. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
82. Grant, C. & Bailey, T. XSTREME: Comprehensive motif analysis of biological sequence datasets. *bioRxiv* (2021) doi:10.1101/2021.09.02.458722.
83. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
84. Lew-Smith, J., Binkley, J. & Sherlock, G. The Candida Genome Database: annotation and visualization updates. *Genetics* **229**, iyaf001 (2025).

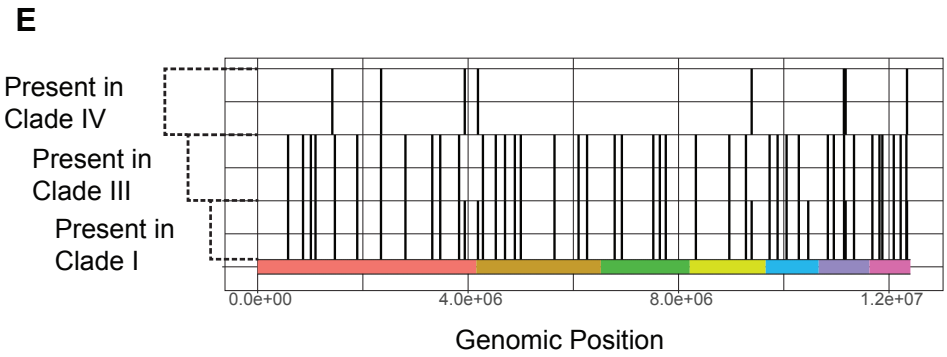
85. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

86. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

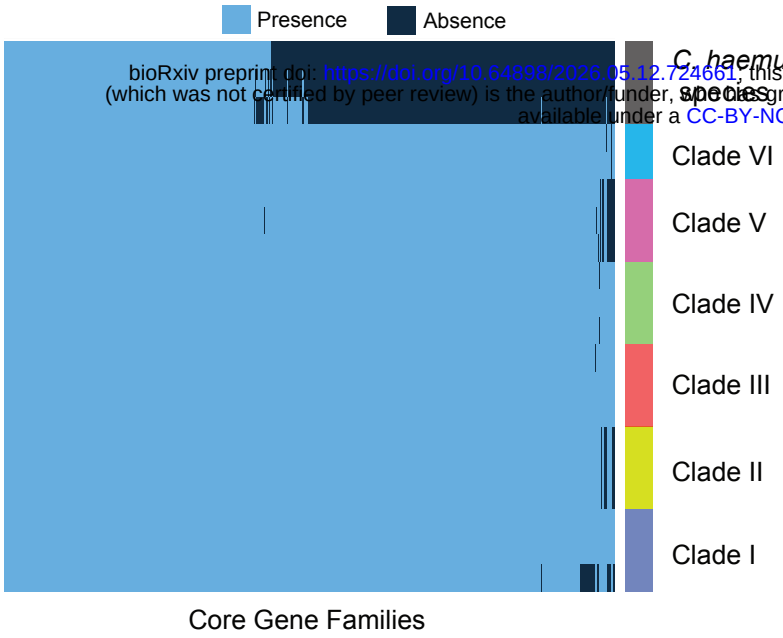


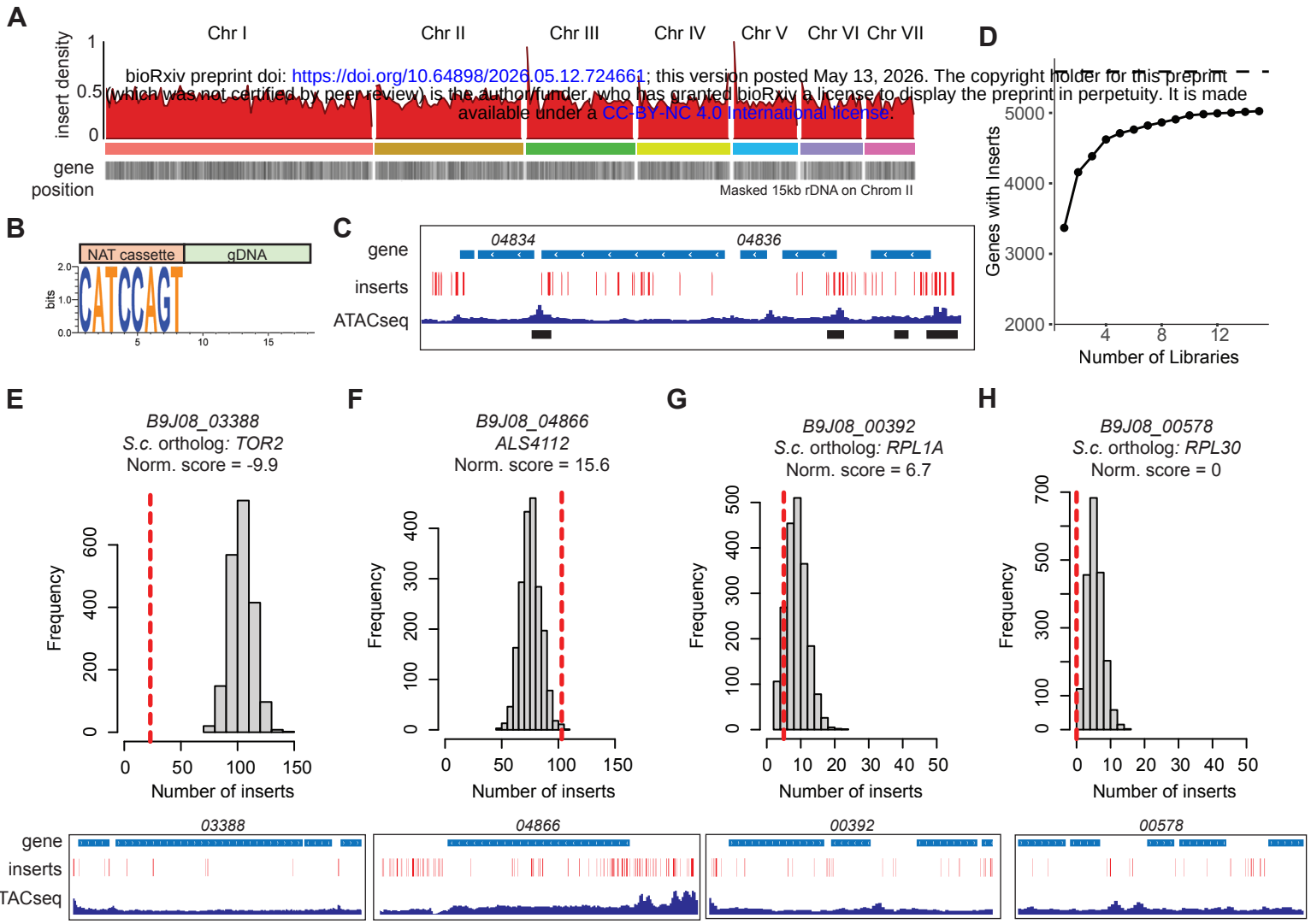


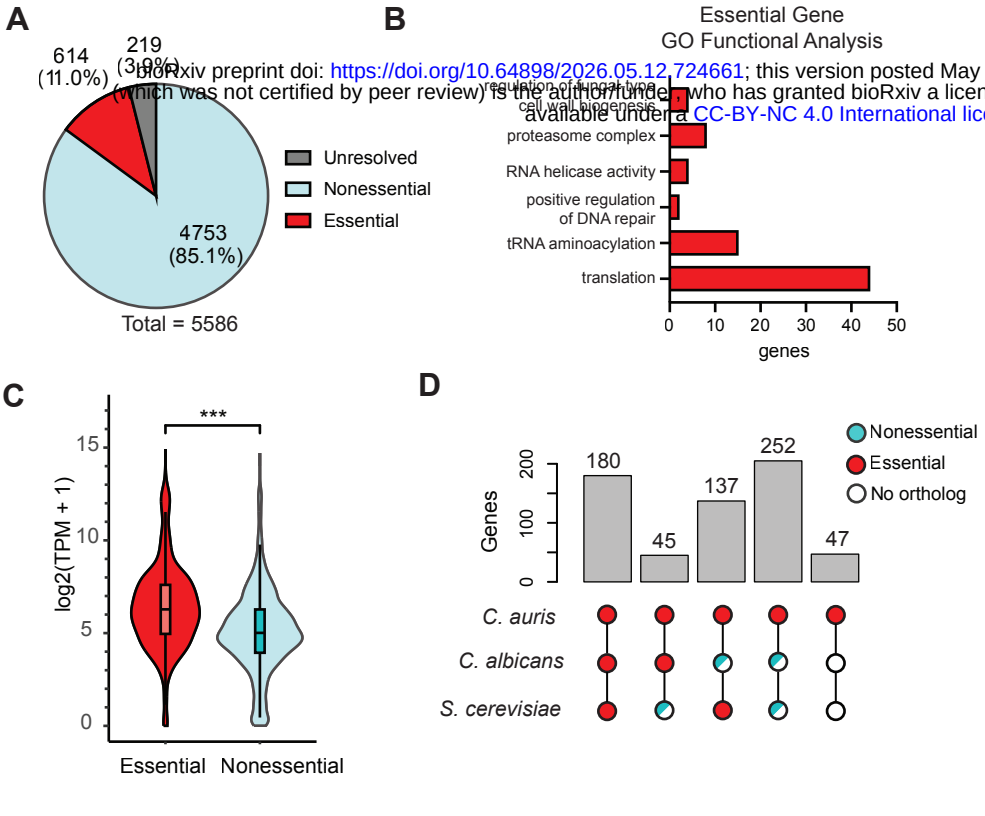


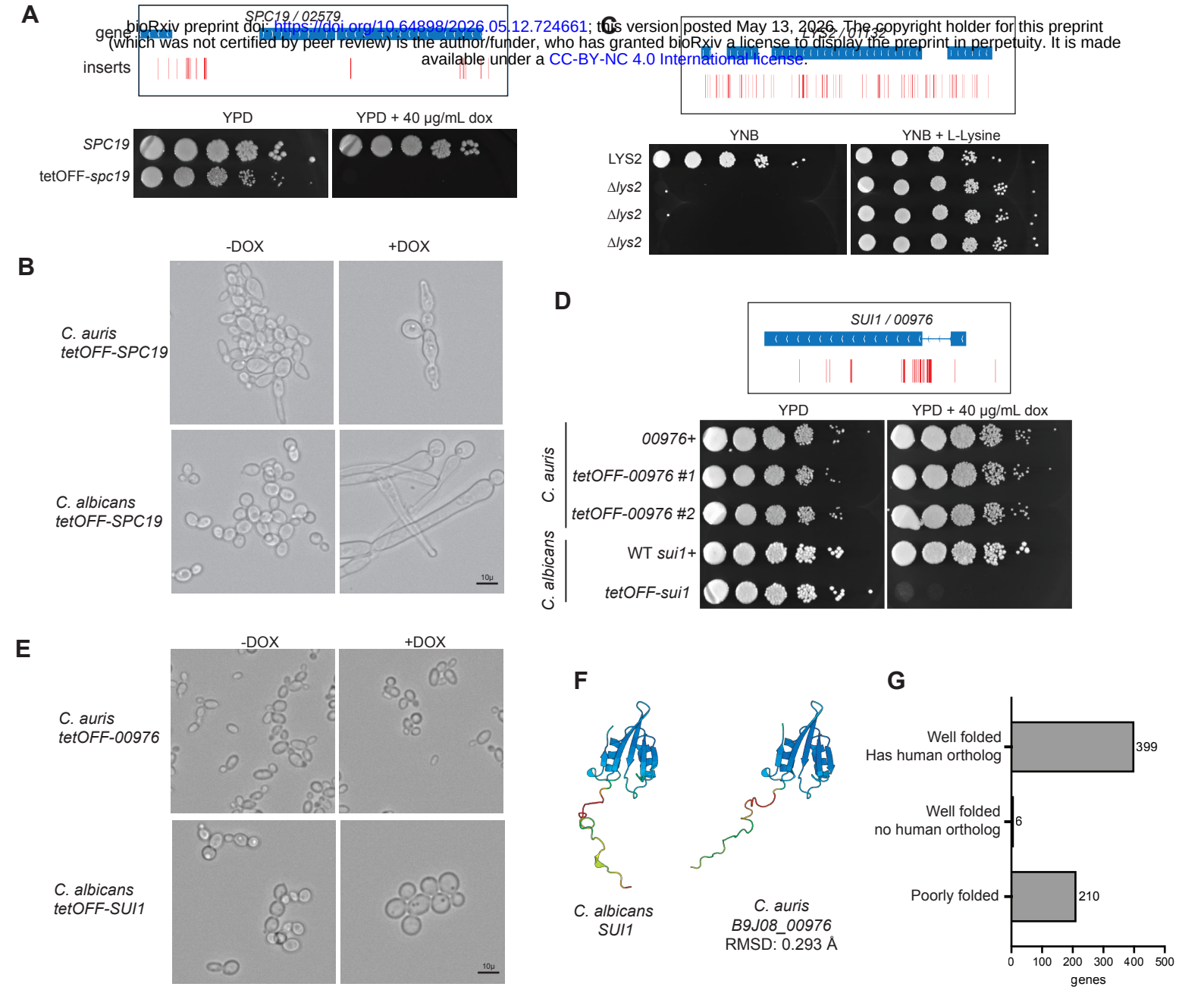


A

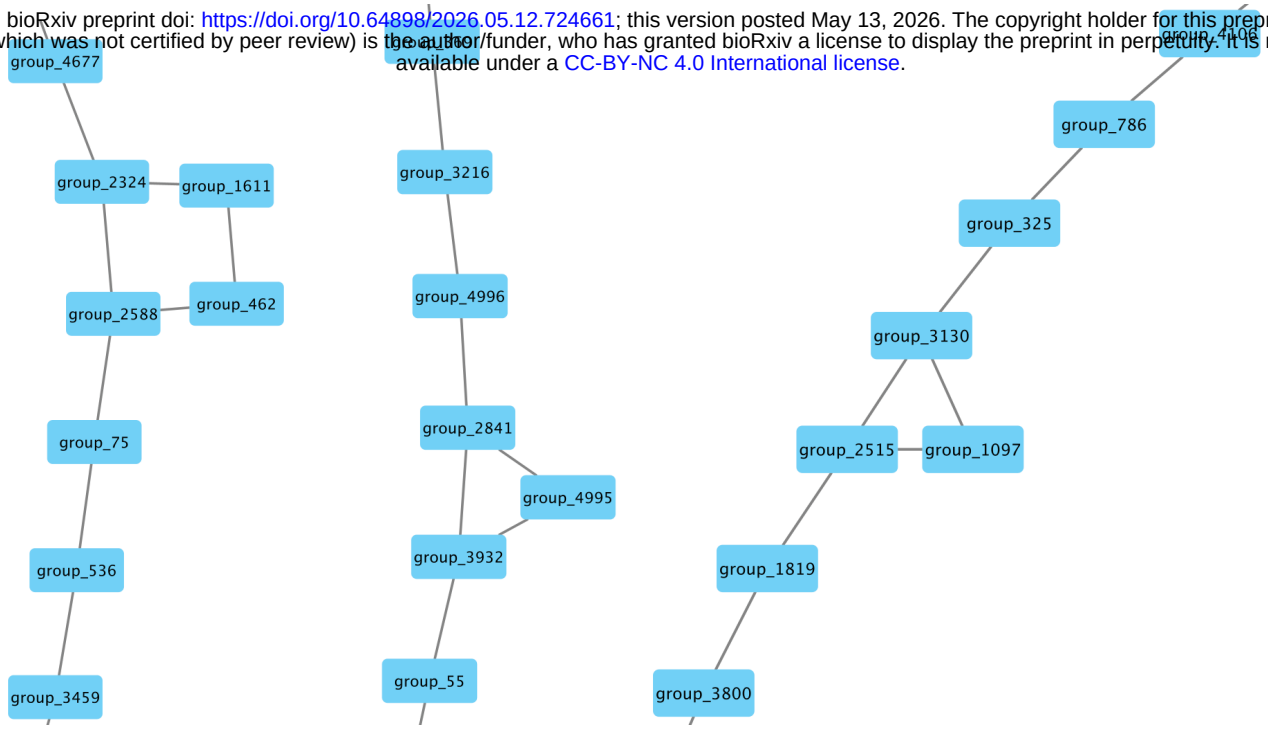


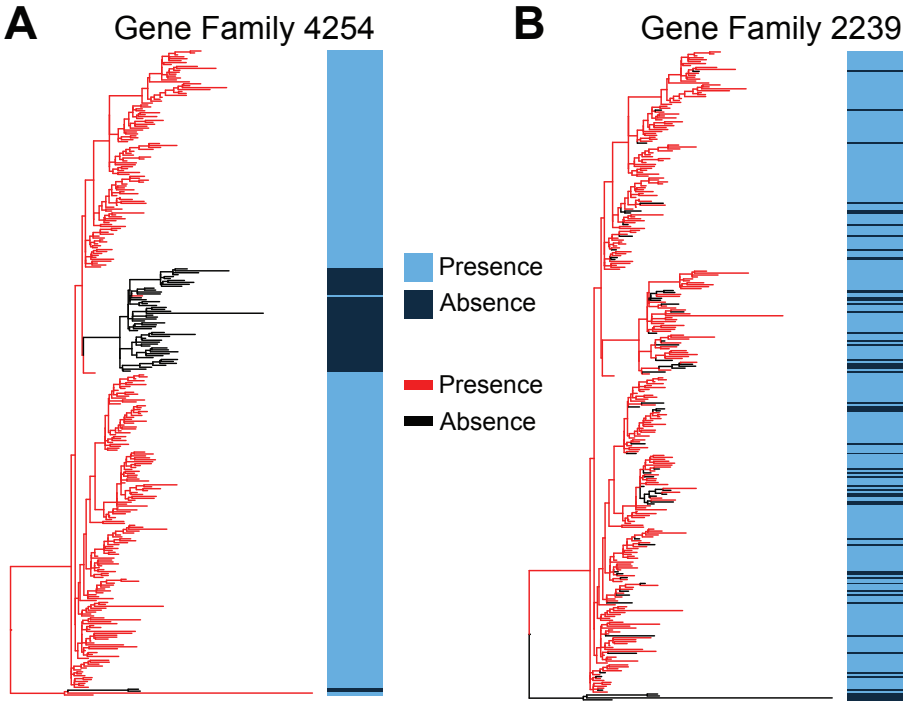




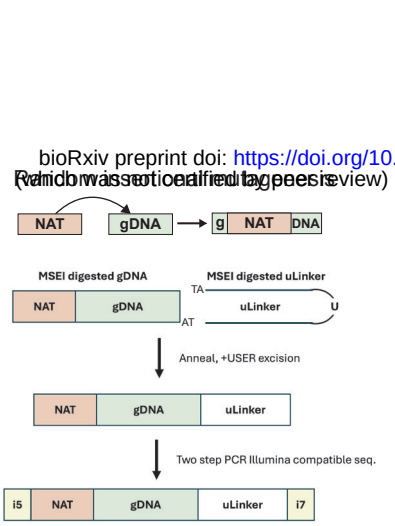


A

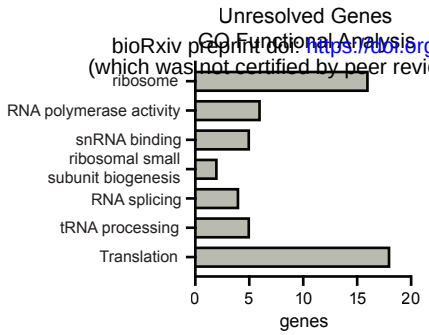




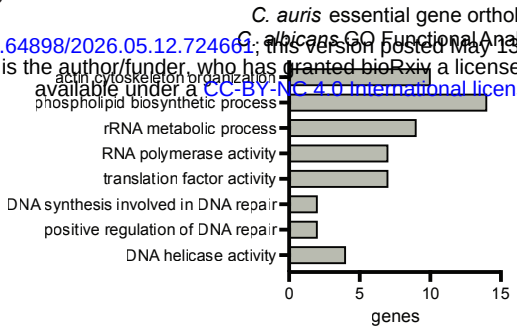
A



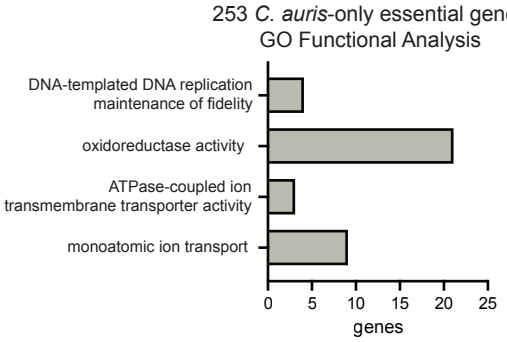
A



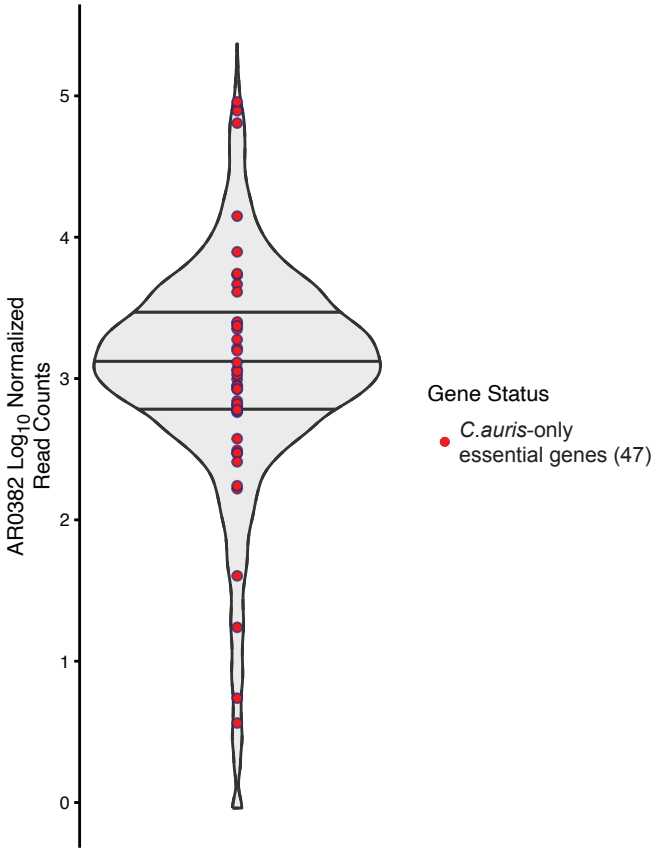
B



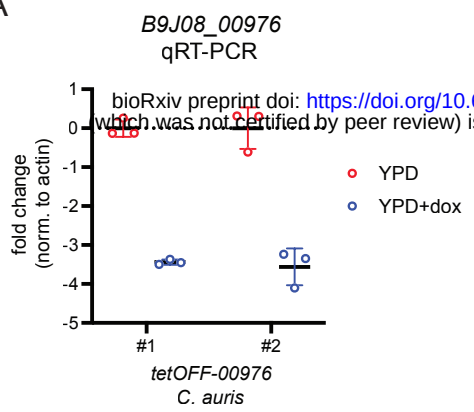
C



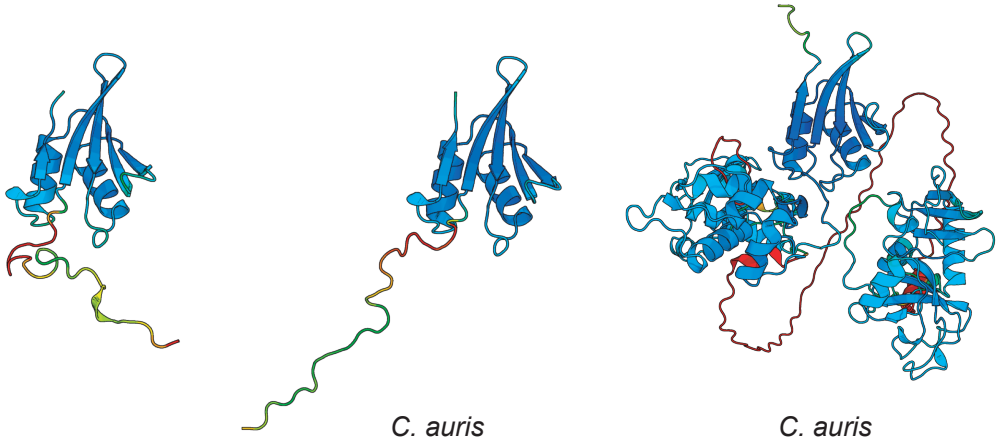
D



A



B



C. albicans
SUI1

C. auris
B9J08_00976
RMSD: 0.293 Å
(85 atoms)

C. auris
B9J08_00176
RMSD: 0.838 Å
(44 atoms)