Long non-coding RNAs (lncRNAs) constitute a poorly studied class of transcripts with emerging roles in key cellular processes. Despite efforts to characterize lncRNAs across a wide range of species, these molecules remain largely unexplored in most eukaryotic microbes, including yeast pathogens of the *Candida* clade. Here, we analyze thousands of publicly available sequencing datasets to infer and characterize the lncRNA repertoires of five major *Candida* pathogens: *Candida albicans*, *Candida tropicalis*, *Candida parapsilosis*, *Candida auris* and *Candida glabrata*. Our results indicate that genomes of these species encode hundreds of lncRNAs that show levels of evolutionary constraint intermediate between those of intergenic genomic regions and protein-coding genes. Despite their low sequence conservation across the studied species, some lncRNAs are syntenic and are enriched in shared sequence motifs. We find co-expression of lncRNAs with certain protein-coding transcripts, hinting at potential functional associations. Finally, we identify lncRNAs that are differentially expressed during infection of human epithelial cells for four of the studied species. Our comprehensive bioinformatic analyses of *Candida* lncRNAs pave the way for future functional characterization of these transcripts.
Advances in high-throughput RNA sequencing (RNA-Seq) in the past decade have shown that eukaryotic cells express abundant and numerous types of non-coding transcripts. One major type of non-coding transcripts is long non-coding RNAs (lncRNAs), which are broadly defined as transcripts longer than 200 bp which do not code for proteins. These molecules have several peculiarities that are consistently reported across a wide range of taxa, including mammals, insects, and plants. Firstly, as compared to protein-coding genes they are generally expressed at lower levels and with a higher cell type specificity. Secondly, they are poorly conserved at the sequence level, and show a rapid evolutionary turnover, being often species-specific. It has been suggested that the functionality of lncRNAs can be attributed to their secondary structure, which can be maintained by selective pressures. However, whether lncRNAs are highly structured is still debated. LncRNAs have been shown to play important roles in numerous processes such as, among others, gene expression regulation, imprinting, splicing, and cell cycle. However, despite extensive research, only a limited number of lncRNAs has been functionally characterized, such as Xist, HOTAIR, Malat1, NORAD, ASCO, among others (see for an extensive review).

Non-coding transcripts are also abundant in fungi where they have been shown to regulate various processes including cell wall remodeling, transcriptional control, and response to nutrients. Most of our knowledge about fungal lncRNAs comes from the model organisms Saccharomyces cerevisiae and Schizosaccharomyces pombe but research on lncRNAs has recently expanded to other fungi, such as Neurospora crassa. Fusarium graminearum, Metarhizium roreri, Pichia pastoris, the white-rot fungus Ganoderma lucidum, and the brown-rot fungi Coniophora puteana and Serpula lacrymans. In accordance with findings for other taxa, studies on fungi consistently show that lncRNAs are generally shorter, have lower levels of expression and GC content, as compared to protein-coding genes. Some of these studies suggested novel putative functional implications of lncRNAs. For example, one study identified 1081 lncRNA transcripts in the filamentous fungus M. robertsi that were differentially expressed upon heat shock, hinting at a potential implication in thermal stress response. In F. graminearum the expression of numerous lncRNAs was shown to be regulated in a stage-specific manner during fruiting body development.

A very limited number of studies, however, have been performed to characterize the role of lncRNAs in fungal virulence. Wang and colleagues investigated these transcripts in the insect pathogen Cordyceps militaris, where >4000 lncRNAs were shown to be regulated during development. Moreover, when xnr1, the final gene of the nonsense-mediated decay pathway, determining the fate of lncRNAs, was knocked-out, an attenuation of virulence and growth rate was observed. The role of lncRNAs in fungal pathogenicity towards humans has been studied in the basidiomycete Cryptococcus neoformans. Chacko et al. identified a genomic locus called RZE1 and found that deletion of this locus results in non-filamentous phenotypes, while its reintroduction restores filamentation. Using site-directed mutagenesis to alter potential translation start codons of RZE1, the authors have shown that this locus can control filamentation even with disrupted putative start codons, indicating that RZE1 functions as an lncRNA, which was also supported by very low sequence conservation of this region across the Cryptococcus clade. Further experiments indicated that RZE1 functions within the nucleus and supposedly indirectly modulates the expression and export of the Znf2 transcription factor which controls morphotype transition in C. neoformans. Interestingly, RZE1 and ZNF2 have proximal genomic positions, and the location of RZE1 rather than its expression per se had a stronger effect on the expression of ZNF2. This was the first study providing evidence of the involvement of lncRNAs in fungal pathogenicity. However, the possible implications of lncRNAs in fungal virulence in other major fungal pathogens remain unknown.

Yeasts from the non-monophyletic Candida genus are among the most widespread human fungal opportunistic pathogens. Up to 30 distinct, phylogenetically diverse, Candida species have been reported to infect humans, mainly when immunocompromised. Candida infections represent a high burden for global healthcare. They range from common superficial infections, such as vaginal candidiasis affecting 75% of the female population, to life-threatening invasive infection, with mortality rates reaching 70%.

Results and discussion

Inference and characterization of lncRNA catalogs in Candida spp. To infer and characterize lncRNA catalogs of the five major Candida pathogens, we used 2645 samples (see Supplementary Data 1 for the full list of samples with corresponding experimental and quality control information) comprising all publicly available RNA-Seq sequencing libraries for these species (Broad dataset, “B”) and RNA-Seq data from a large-scale in vitro host-pathogen interaction study of the four species C. albicans, C. tropicalis, C. parapsilosis, and C. glabrata with human vaginal epithelial cells (Specific dataset, “S”, see Materials and Methods for details). We performed genome-guided transcriptome assemblies for each individual species and for each of B and S datasets independently. Assembled transcriptomes from B and S datasets were merged for each species to produce a final catalog of predicted lncRNAs (Fig. 1, see “Materials and Methods” for details).

Based on their genomic coordinates, we classified lncRNAs into intergenic (“i”), i.e., transcripts that do not overlap with protein-coding genes or other features, and antisense (“a”), i.e., lncRNAs overlapping coding genes or other features on the opposite DNA strand. Full lncRNA catalogs for all species are available in Supplementary Data 2 and are deposited in Candidamine—an integrative data warehouse for Candida yeasts available at https://candidamine.org/.
Prior to investigating the lncRNA repertoires in more detail, we assessed the impact of the chosen data analysis pipeline (i.e., genome-guided assembly by Stringtie) on the results of transcriptome assembly, which is the main prerequisite for robust lncRNA identification. To this end, we compared the transcriptome assemblies with those produced by an alternative de novo transcriptome assembly approach as implemented in Trinity. We observed that both assemblers largely produced consistent results in terms of reconstructing the reference annotated features, although the genome-guided method employed here rendered assemblies with higher specificity, sensitivity, and lower fragmentation rate (see the results of these analyses in Supplementary Note 1). Although the reference annotated features used in the benchmark do not include lncRNAs, we consider that it is reasonable to expect a better performance of Stringtie across transcript classes. Therefore, given the overall better performance of Stringtie in reconstructing reference transcripts, the results of the genome-guided transcriptome assembly approach were used in all downstream analyses.

The largest number of lncRNAs (Fig. 2a) was detected in C. albicans, with 5763 antisense and 1459 intergenic transcripts, followed by C. auris (4759 and 839, respectively), C. parapsilosis (3038 and 1499), C. tropicalis (987 and 1568), and C. glabrata (989 and 449). For species of the CTG clade (i.e., all except C. albicans), differences in the number of lncRNAs are mainly driven by the large number of antisense transcripts found in C. albicans (5763) and C. auris (4759) as compared to C. parapsilosis (3038) and C. tropicalis (987).

Of note, such drastic differences are not observed for C. albicans when analyzing only the samples of the host-pathogen interaction study (S dataset, 1088, 627, 968 antisense lncRNAs in C. albicans, C. tropicalis, and C. parapsilosis, respectively). Both

Fig. 1 Schematic representation of the bioinformatics workflow for lncRNA prediction and analysis. See description in the text for details.
biological differences and differences in the diversity of analyzed samples (Supplementary Data 1 and Figure S1) can potentially influence the number of lncRNAs detected in our study. An additional parameter that can have a strong effect on the amount of identified lncRNAs is the number of analyzed samples for our final transcriptome assemblies, which was largest for C. albicans (n = 699), followed by C. parapsilosis (n = 86), C. auris (n = 61), C. tropicalis (n = 53), and C. glabrata (n = 51). To assess the impact of the difference in the number of analyzed samples, we repeated the lncRNA predictions using sample subsets and produced saturation plots showing the dependency of the number of predicted lncRNAs (Supplementary Fig. S2). This analysis showed that the number of antisense lncRNAs in C. albicans reaches a plateau for subsets of ~200 samples or larger. This suggests that few novel antisense lncRNAs in C. albicans may remain to be discovered. For the other species, considering they have at most 86 samples (for C. parapsilosis), it is likely that their antisense lncRNA catalogs are not complete, and might be expanded with new datasets. However, additional biological and technical factors that might influence this result include pervasive transcription, transcriptional noise, and varying efficiency of strand-specific library preparation protocols. On the other hand, extensive antisense transcription has been previously observed in C. albicans and S. cerevisiae, although to a lesser extent than observed in our study. Additionally, it must be noted that our analyses are based on the use of reference genomes and therefore it is unclear how genomic variability across strains of the same species may impact lncRNA predictions. However, this problem is inherent to any reference-based inference, and our approach can be extended to multiple (i.e., clade-specific) reference genomes.

While the number of antisense lncRNAs varied significantly across the four species of the CTG clade, the number of intergenic lncRNA was similar across the three more closely related representatives of this clade (1457–1581), with C. auris having a somewhat lower number (842). In contrast, the distantly related C. glabrata has a much lower number of intergenic lncRNAs compared to all other species (444). A lower number of intergenic lncRNAs in C. glabrata is also apparent when restricting the analysis to the S dataset (267, compared to 1223, 943 and 794 for C. albicans, C. tropicalis, and C. parapsilosis, respectively). Thus, the considerably lower number of intergenic lncRNAs in C. glabrata may reflect a true biological difference, in accordance with the large phylogenetic distance between C. glabrata and the other considered species. Finally, saturation plots for intergenic lncRNA reached a plateau with only 30–40 samples, suggesting that the obtained catalogs are comprehensive. Of note, a decay in the total number of intergenic lncRNAs is observed in C. albicans as more datasets are analyzed, likely resulting from the fusion of previously fragmented transcripts.

Further, we assessed the distribution of lncRNAs along chromosomes (Supplementary Figs. S3 and S4). For C. albicans, C. parapsilosis, and C. auris, antisense transcripts were relatively evenly distributed across the chromosomes, as compared to more variable distribution in C. tropicalis and C. glabrata.
(Supplementary Fig. S3). Intergenic transcripts were unevenly distributed in all species, with some apparent hotspot regions. For example, in chromosomes 1A and 2A of *C. albicans*, these hotspots are observed near centromeres. In contrast, *C. glabrata* shows high accumulation of intergenic lncRNAs at the terminal (telomeric) sites of some chromosomes, such as chromosomes D, E, F, H, and J. This pattern of lncRNA hotspots (both intergenic and antisense) in *C. glabrata*, but not in other species was also clearly observed when calculating the number of lncRNAs relative to the distance to the closest telomere across all chromosomes (Supplementary Fig. S4).

We then assessed several major features of lncRNAs and compared them to those of protein-coding genes and, where available, previously annotated lncRNAs. Consistent with studies in other organisms, we found that, compared to protein-coding genes, both lncRNA types tend to be shorter, have lower GC content (albeit higher than in intergenic regions), and lower levels of expression (Fig. 2b, see Supplementary Data 3 for the mean values of all parameters across species and comparisons).

We further used population-scale variant calling data of each species derived from 1976 publicly available genome sequencing samples (see “Materials and Methods” for details) to analyze the variability of lncRNAs and compare it to protein-coding genes and intergenic regions. The analysis of genetic variants (Fig. 2b) showed that lncRNAs accumulate more variants than protein-coding genes but significantly fewer than intergenic regions (with an exception of *C. auris* where intergenic lncRNAs had higher variability than intergenic regions, see Supplementary Data 3), indicating that lncRNAs are evolutionarily constrained.

We also performed a similar set of analyses for 17 previously predicted lncRNAs in *C. glabrata* and, observed that these transcripts were generally more similar to protein-coding genes than to lncRNAs found in our study (Supplementary Fig. S5). This might be explained either by the small sample size of previously annotated lncRNAs, or their possible misidentification as noncoding transcripts.

It has been previously shown that lncRNAs in a wide range of taxa can encompass transposable elements (TEs) and genomic repeats, which can influence their origin, architecture, evolutionary trajectories, and regulation. We, therefore, analyzed whether and to which extent lncRNAs in *C. albicans* harbor these elements and compared it to protein-coding genes (Supplementary Fig. S6). This analysis (see “Materials and Methods” for details) showed that many lncRNAs, both intergenic and antisense ones, harbor repeat regions in all studied species, although to a different extent. The highest values were observed in *C. tropicalis* (50.5% of intergenic and 50.9% of antisense lncRNAs) and *C. albicans* (48.5% of intergenic and 50.8% antisense lncRNAs) and were decreasing following the phylogenetic relationships of the species, reaching in *C. glabrata* to 20.7% of both intergenic and antisense lncRNAs. By far the most prevalent type of repeats found within lncRNAs are simple repeats, followed by low complexity repeats, while the prevalence of LINE, SINE, and LTR repeats was generally low. Importantly, for all species except *C. glabrata*, we also observed a high proportion of protein-coding genes overlapping repeats, which however was lower than in the case of lncRNAs. For example, for *C. albicans* 40.1% of protein-coding genes overlapped repeats. Of note, this pattern of having more repeat regions in lncRNAs than in protein-coding genes has been reported in a wide range of taxa.

Classification of lncRNAs into conserved families. We next explored the evolutionary relationships of lncRNAs across the studied species. Considering that lncRNAs generally show low levels of sequence conservation and may adopt conserved secondary structures, we used three alternative approaches to establish their potential relatedness: sequence similarity, structural similarity, and synteny. This is specifically relevant here, as the studied *Candida* species are phylogenetically very diverse. As antisense lncRNAs partially overlap protein-coding genes, we excluded them from this analysis. We first identified one-to-one best reciprocal BLAST hits between each pair of species (see Materials and Methods for details) and used a previously developed clustering methodology to unify all pairwise species comparisons and define lncRNA families across species. As expected, this analysis identified a very small number of conserved lncRNA families—merely one between *C. tropicalis* and *C. parapsilosis* (see Supplementary Fig. S7) containing one lncRNA from each species (MSTRG.4971.1 and MSTRG.364.1, respectively). To validate this result with a different clustering approach, we ran OrthoMCL analysis using Synima software which rendered the same lncRNA family as obtained by BLAST reciprocal hits. These results highlight the overall low sequence conservation between lncRNAs, and the high levels of species divergence.

We then analyzed structural similarities of lncRNAs using the Beagle software which performs pairwise alignments of secondary structures (see “Materials and Methods” for details). In this approach the majority of lncRNAs were classified into structural families (~98.2%, Supplementary Fig. S7) and most families (84.13%) were shared by all five species, which is in stark contrast with BLAST results. It was unclear, however, whether this similarity is the result of higher evolutionary constraints in the structure, or whether it can be attributed to a low specificity of the structural comparison approach. To test this, we repeated the above mentioned analysis 50 times using randomly reshuffled sequences of lncRNAs, and observed similarly large fraction of lncRNAs classified into families (mean = 97.11%, p = 0.86 compared to the real rate), indicating that the alignment of secondary structures with Beagle has low specificity and does not allow drawing evolutionary inferences.

We finally searched for syntenic lncRNAs using a methodology developed and validated in (see “Materials and Methods” for details). This analysis revealed a significantly higher number of evolutionary related lncRNAs (n = 881, 26.8%) than the BLAST-based approach, but considerably lower than in the case of secondary structure analysis (Fig. 3, see Supplementary Data 4 for pairwise synteny information across species). We repeated this analysis with varying levels of stringency of synteny classification parameters and consistently obtained similar results (Supplementary Data 5 and 6). As expected from their closer evolutionary relationships, the majority of the syntenic relationships were observed between species of the CTG clade, while the level of synteny between *C. glabrata* and other species was lower.

We further inspected syntenic families for the presence of conserved sequence motifs shared by transcripts within a family, which could additionally support functional or evolutionary relatedness. Motif discovery analysis using MEME detected 91 families (or 21.5%) with shared conserved motifs. This fraction of lncRNA families with shared motifs was significantly higher than random expectation (p < 0.05) as assessed in a set of randomly generated families (see “Materials and Methods”) which on average achieved a 0.74% rate of motif discovery. A similar analysis using an independent graph-based approach implemented in lncLOOM showed that 85 families (20%) had shared sequence motifs. Additionally, the same analysis using random families showed a significantly lower rate (8.5%, p < 0.05). We then compared the results of the two approaches and found 33 lncRNA families which were identified to have conserved
sequence motifs by both tools. Further analysis indicated that despite the significant difference between computational methodologies employed by MEME and lncLOOM, there were 5 lncRNA families where motifs found by the two software were overlapping.

Overall, our results suggest that despite large evolutionary distances between Candida species and the overall poor sequence conservation, these pathogens share syntenic lncRNAs which possess short patches of conserved sequence motifs. Such short patches of conserved sequences in lncRNAs have been broadly reported (see 15, 80 for recent reviews), inspiring the “RNA modular code” hypothesis81, which posits that these elements can form discrete functional secondary and tertiary structures mediating interactions with other molecules. Testing this requires a combination of complex experimental and bioinformatics analyses. In this context, our findings of syntenic lncRNA families with conserved sequence motifs in major Candida pathogen open novel opportunities for disentangling those relationships in a targeted manner for future research.

Co-expression analysis. The potential functions of the newly identified lncRNAs are unknown, but their patterns of co-expression with protein-coding genes of known function can hint on possible roles82,83. Considering this, we carried out a gene co-expression network analysis for intergenic lncRNAs and protein-coding genes using the WGCNA approach (see “Materials and Methods”). After inspecting principal component analysis plots (Supplementary Fig. S1) to remove outliers and filtering out lowly expressed genes (TPM < 0.1 in more than 80% of samples), we obtained sufficient power values ($\beta = 12–16$) to generate scale-free co-expression networks for each of the species (Supplementary Fig. S8).

For all species, a co-expression network analysis identified multiple highly interconnected gene modules ($n = 9–20$, Supplementary Fig. S9). Interestingly, lncRNAs were present in the majority of modules, with the only exception of C. albicans, indicating that lncRNAs are common members of co-expressed gene clusters. Despite being widely distributed across networks, lncRNAs have significantly lower co-expression connectivities than protein-coding genes (Wilcoxon rank sum test $p$ value $< 0.05$). Despite this generality, lncRNAs MSTRG.7139.1 of C. tropicalis and MSTRG.4801.1 of C. glabrata were the most highly connected nodes (i.e., the hubs) in their corresponding modules (Supplementary Fig. S7).

To gain functional insights on the lncRNAs involved in modules, we performed enrichment analyses of GO terms, KEGG pathways, and PFAM domains of the protein coding genes in all
modules (see Supplementary Data 7 for all identified enrichments of each module and species). We observed a wide variety of enrichments across all the modules, including terms related to fungal pathogenicity, such as "adhesion of symbiont to host" and "pathogenesis" for C. albicans (module "darkred") and "filamentous growth" for C. tropicalis (module "coral1").

Identification of lncRNAs specifically regulated during epithelial infection. To investigate in more detail the possible implication of Candida lncRNAs in virulence, we resorted to the S dataset, which comprises host-pathogen expression data through the time-course of epithelial infection of four of the studied pathogens—C. albicans, C. tropicalis, C. parapsilosis, and C. glabrata. To this end we performed a differential expression analysis to identify lncRNAs (both intergenic and antisense) which are differentially expressed throughout the infection (see Supplementary Data 8–11 for data for C. albicans, C. tropicalis, C. parapsilosis, and C. glabrata, respectively). As shown in Fig. 4a, there are numerous lncRNAs that significantly change their expression upon interaction with epithelial cells from the initial time point of infection, with the number of differentially expressed lncRNAs increasing as the infection progresses. The 24 h control samples present in the S dataset allowed us to identify lncRNAs which are differentially expressed in the absence of human epithelial cells, accounting for the effect of time and changes in the growth medium. We observed that the process of infection and normal growth in culture medium deregulate largely overlapping sets of lncRNAs (Fig. 4b). This phenomenon has been already observed for protein-coding genes of these species85,84, indicating that most non-coding and coding genes related to pathogenesis of these species are also related to the standard growth metabolism. We nevertheless identified a substantial number of infection-specific lncRNAs, which are differentially expressed exclusively due to the infection process (n = 191–270, depending on the Candida species).

Further analysis of the "infection-specific" lncRNAs showed no significant differences between their network connectivities compared with other lncRNAs. Additionally, we observed that a large portion of "infection-specific" intergenic lncRNAs (~10–50%) are involved in modules. Of note, modules containing infection-specific lncRNAs had various enrichments (see Supplementary Data 7), some of which were directly related to pathogenicity, such as "pathogenesis" and "adhesion of symbiont to host" for C. albicans (module "darkred") and "filamentous growth" for C. albicans (module "darkred") and C. tropicalis (module "coral1") (Fig. 5). To ensure that the co-expression signals of "infection-specific" lncRNAs in these modules are not spurious, we investigated the expression levels of the "infection-specific" lncRNAs with the highest WGCNA weight and corresponding co-expressed coding genes across the analyzed samples, which reinforced their co-expression patterns (Supplementary Fig. S10). Interestingly, this analysis also showed that, in C. tropicalis "coral1" module, the protein coding gene which is co-expressed with the two infection-specific lncRNAs with highest weights (i.e., MSTRG.6541.1 and MSTRG.6542.1) is an infection-specific protein coding gene CTRG_00938 (as identified by Pekmezovic et al.85). This gene has an unknown function in C. tropicalis, but has an ortholog in C. albicans (CR_01500W) which is implicated in filamentous growth (according to Candida Genome Database85).

We then assessed syntenic relationships of "infection-specific" lncRNAs across species (Supplementary Fig. S11), which showed the presence of a few syntenic families sharing infection-specific lncRNAs between closely related species of CTG clade. Namely, there were three families between C. albicans, C. tropicalis, and C. glabrata.

---

Fig. 4 lncRNAs during fungal infection of epithelial cells. a lncRNAs expression dynamics plots based on log2 fold changes compared to time point 0. Each line (up-regulated—red, down-regulated—blue) corresponds to the fold change of expression levels of a lncRNA. Numbers on the plots indicate the number of differentially expressed lncRNAs ([log2 fold change ] >1.5, padj < 0.01, in red—intergenic lncRNAs, in black—antisense lncRNAs); b Venn diagrams of differentially expressed (DE*) lncRNAs during infection (in green) and in control samples (in blue). To identify infection-specific genes with a higher stringency, we applied filters of [log2 fold change ] >0 and padj < 0.01 to the DE lncRNA sets in both conditions in order to maximize the overlap between them, hence ensuring that lncRNAs in control samples even with small fold-changes are discarded. The numbers of "infection-specific" lncRNAs are indicated in green-only portions of Venn diagrams.
one family *C. albicans* and *C. parapsilosis* and two families between *C. tropicalis* and *C. parapsilosis*. Overall, this analysis suggests that infection-specific genes tend to be specific for each *Candida* pathogen, in agreement with previous results for protein coding genes65.

**Concluding remarks.** Despite intensive research on lncRNAs during the recent decade86, these enigmatic transcripts have never been systematically investigated in human fungal pathogens from the *Candida* clade. Here, we mined and analyzed thousands RNA-Seq samples available for the studied yeasts, and reported comprehensive lncRNA catalogs of the five major *Candida* pathogens. As in other eukaryotic species, lncRNAs are abundant in these *Candida* yeasts. We classified the identified lncRNAs into intergenic and antisense transcripts, and show that antisense transcription of lncRNAs is widespread in *Candida*. These catalogs constitute a valuable resource to guide further research into the role of lncRNAs in these organisms. Incidentally, one of our predicted lncRNAs in *C. auris*, namely MSTRG.10503.1, was experimentally identified and characterized in an independent study87 while our manuscript was in revision. This lncRNA, named DINOR in that study, was shown to act as a virulence regulator.
factor and a regulator of various stress responses in C. auris. This independent confirmation of one of our predicted lncRNAs attests for the validity of our approach for identifying biologically relevant lncRNAs.

We found that Candida lncRNAs exhibit similar general properties as lncRNAs in other species, i.e., they are shorter, and have lower GC content, expression levels and evolutionary constraints when compared to protein-coding genes. From the evolutionary standpoint, lncRNAs in Candida show poor primary sequence conservation, but importantly they accumulate less sequence variation than intergenic regions, which hints to the existence of selective pressures acting on these transcripts. Moreover, lncRNAs showed detectable levels of synteny between CTG clade members exclusively due to interaction with the human host, and were co-expressed with protein coding genes.

Moreover, lncRNAs showed detectable levels of synteny between CTG clade members re- and C. glabrata, representing strand-specific sequencing libraries with 2 × 50 and 2 × 75 bp read length. The samples which had traces of adapter sequences and/or poor quality bases were trimmed with Trimmomatic using <ADAPTERS.fa>:2:30:10 LEADING:1 TRAILING:1 SLIDINGWINDOW:4:1 MINLEN:<5> command. Subsequently, since most of the S dataset libraries comprised dual RNA-Seq data (i.e., mixed human and fungal RNA), we mapped the data to the concatenated reference genome of each fungus and human. Human genome GRCh38 and annotations were obtained from Ensembl database (last accessed on 9 August 2017). Then, we generated fungus-only bam files by subset mapping fungal reads from the pooled bam files using samtools v. 1.3.1.96.

Detailed information about all RNA-Seq samples used in this study is available in Supplementary Data 1.

Computational prediction of lncRNAs. For each sample we performed genome-guided transcriptome assembly using Stringtie v. 1.3.396. We then compared the result of Stringtie-based genome-guided transcriptome assembly with de novo transcriptome assemblies reconstructed by Trinity v. 2.8.5 software97, using the available reference annotations from CGD as gold standard. This comparison showed overall consistent results between the two softwares, and revealed that genome-guided Stringtie assembly resulted in higher specificity, sensitivity and precision of all transcriptome re-assembly rates (see Supplementary Table S1 for detailed description of the performed comparisons). For this reason, all downstream analyses for lncRNA identification were performed using Stringtie results.

For further, each species we merged all assembly gtf files produced by Stringtie from each sample using Stringtie merge with -g option which resulted in a unified transcriptome. These transcriptome annotations were compared with the original genome annotations of each species using gffcompare v. 0.11.2 to identify novel transcripts.

For C. albicans, our initial mapping and assembly strategy resulted in transcripts with artificially long introns spanning several hundred thousand bases. To avoid this, we repeated the analysis for all species setting the TopHat2 option --max-intron-length to 1000, which corresponds to approximate maximum intron length in fungal genomes97. Additionally, for C. albicans we removed four assemblies from the project PRJNA294249 (assembled from SRR2153488-SRR2153491) using <ADAPTERS.fa>:2:30:10 LEADING:1 TRAILING:1 SLIDINGWINDOW:4:1 MINLEN:<5> command. The remaining data, i.e., 666 libraries for C. albicans, 129 for C. parapsilosis, 125 for C. glabrata, 86 for C. auris, and 46 for C. tropicalis, was phased, in our analysis we used only haplotype A for read mapping to C. albicans genome. The fungus-only bam files by subsetting mapped fungal reads from the pooled bam files using samtools v. 1.3.1.96.

Overall expression levels of lncRNAs. To assess the overall expression levels of lncRNAs and compare them to those of protein-coding genes, we calculated the
read counts of both transcript categories in all analyzed samples using FeatureCounts v. 1.6.135. The count data were normalized by transcript length and library size, resulting in transcripts per million (TPM) values.

**Analysis of sequence variation.** To assess the variability of IncRNA sequences and compare it with that of protein-coding genes and intergenic regions, we utilized variant calling data of the studied species available in the Candamidine database at https://candamidine.org/, which comprises variants obtained from all 1976 publicly available DNA sequencing data of these yeasts. In particular, we used variants from 652 samples of *C. albicans*, 420 samples of *C. glabrata* and 51 samples of *C. parapsilosis*, 764 samples of *C. auris*, and 89 samples of *C. tropicalis*. Using bedtools v.2.29.2135 with intersect function we calculated the number of variants located in both types of IncRNAs, protein coding genes, and intergenic regions across the studied species. Intergenic regions were retrieved using bedtools complement function applied to genome annotation files obtained after gffcompare.

**Analysis of genomic repeats and TE elements.** Repeat calling was performed using perSVade pipeline v. 0.10 (https://github.com/Gabaldonlab/perSVade), which runs RepeatModeler v. 2.0.1134 and RepeatMasker v. 4.0.9 (Smit, AFA, Hubley, R & Green, P. RepeatMasker Open–4.0. 2013-2015, http://www.repeatmasker.org) (both with default parameters) on the query genomes, and reports a table of found repeat regions. For this study we analyzed the following types of repeats—simple repeats, low-complexity repeats, LTRs, LINE/SINE and repeats classified as unknown by Repeatmodeler/Repeatmasker. We then used bedtools intersect to calculate the proportion of each IncRNA type and protein coding genes overlapping different classes of found repeat regions, requiring at least 50% of a given repeat’s sequence to overlap IncRNA/protein-coding gene.

**IncRNAs gene family classification.** To assess evolutionary relationships between predicted IncRNAs across species, we used several independent strategies, namely blast reciprocal hits, secondary structure similarity, and analysis of synteny. Considering that overlapping features of antisense IncRNAs can potentially influence the results, only intergenic transcripts were used for these analyses. To define best reciprocal hits between all possible pairs of species we used BLASTn v.2.9.2. To this end, we built a custom BLAST database for the set of IncRNAs of each species. Then, each set of IncRNAs was aligned against each database with BLASTn using cut-off of e-value <1e-3, and -max_hsp 1 and -max_target_seq 1, which selects only the best alignment between matched query-sequence pair. Best reciprocal hits were selected. We also repeated this analysis with parameters -max_hsps 10 and -max_target_seqs 10 and obtained the same results. Additionally, to test a different clustering approach, we have run OrthoMCL analysis139 using Synima pipeline (downloaded on August 1st, 20211). To assess the reliability of the IncRNAs based on their secondary structures, we first used RNAfold v. 2.4.14 from the ViennaRNA package145 to obtain secondary structures for all studied intergenic IncRNAs. Then, using Beagle v. 0.276 with the local alignment mod, the IncRNAs structures of each species were aligned against those of other species in a pairwise manner. Similarly to BLASTn best reciprocal hits, we aligned each IncRNA in one species against the IncRNAs of the other species (at least z Score > 3) and p < 0.01. Reshuffling of IncRNA sequences for testing the specificity of secondary structure alignments was done using fasta-shuffle-letters from MEME suit v. 4.11.2136. We classified intergenic IncRNAs into syntenic transcripts using a methodology developed elsewhere146,147. Briefly, we first observed the information of 1-to-orthologs between protein-coding genes in the four *Candida* species from CGD and CGCDB106. Then, we defined pairwise syntenic relationships between the IncRNAs of the studied species using the *synteny_nematodesv4GH.py* script from18 with modifications directed to the analysis of five instead of four species, and also to correct the detected pairs when transcript identifiers in the study. The script was run with parameters 3 3 1, i.e., considering three protein coding genes at each side of a given IncRNA, a minimum of three shared genes for each pairwise comparison between species, and a minimum of one shared gene at each side of a IncRNA. The analysis identified syntenic IncRNAs between species. We also performed the synteny classification with varying stringency of the above-mentioned parameters (see Supplementary Data 5 and 6).

Finally, to cluster best reciprocal hits, results of secondary structure alignments and pairwise syntenic IncRNAs into IncRNA families across species, we used *classifyFamilies* v. *VennGraph.py* script from18 which was modified for our study. The modified scripts of RNA scripts from18 are available at our GitHub page https://github.com/Gabaldonlab/IncRNAs.

**Analysis of sequence motifs in syntenic families.** We assessed whether the IncRNAs of identified syntenic families shared sequence motifs within a family by employing two independent approaches. First, we scanned the sequences of IncRNAs of syntenic families for motifs using MEME v. 4.11.27. Motifs with e-value < 0.05 were considered as significant. We considered a syntenic family to have a shared motif if all the members of the family contained at least one shared motif. To assess how the identified number of syntenic families with shared motifs was different from random expectation, we simulated the same number and size of IncRNA families by randomly choosing IncRNAs and performed the shared motif discovery as described above. The simulated families comprised transcripts randomly selected from the whole set of intergenic IncRNAs, preserving the number of IncRNAs per species observed in the real syntenic families. We repeated this simulation 100 times and compared the observed rate of shared motifs with the distribution in randomly formed IncRNA families. Second, we used the recently developed tool IncLOOM v. 1.07, which uses a graph-based approach to identify shared sequence motifs in evolutionarily related sets of sequences. For each IncRNA syntenic family we have run IncLOOM with -r 100 option, letting the software to randomly simulate the input sequences to calculate p- and e-values of motif discovery. As in the case of MEME, a syntenic family was considered to have a shared motif if all the members of the family had at least one motif with p- and e-values < 0.05. Additionally, we have run IncLOOM on 100 randomly generated families and compared the results with those of the real dataset.

**Co-expression analysis.** We assessed the patterns of gene co-expression across all intergenic IncRNA transcripts using the weighted correlation network analysis approach implemented in WGCNA v. 1.69. As in case of gene family classification, antisense transcripts were discarded from this analysis. We used log2(TPM + 1) as expression values. For each species, we first selected the β power values using the “pickSoftThreshold” function implying an unsigned network. The minimum β value reaching 80% of scale-free network topology, specifically β = 12 for *C. albicans*, β = 16 for *C. tropicalis*, β = 14 for *C. parapsilosis*, β = 12 for *C. auris*, and β = 12 for *C. glabrata*, was used for downstream analysis. To reach the optimal values of β, we removed samples comprising outliers as identified by inspection of principal component analysis (PCA) plots based on expression values (Supplementary Fig. S1). Namely, we removed samples of the following accession numbers: SRP099169 and SRP083839 for *C. tropicalis*, SRP151798 and SRP041812 for *C. auris* and for *C. glabrata*, was used for downstream analysis. To reach the optimal values of β, we removed samples comprising outliers as identified by inspection of principal component analysis (PCA) plots based on expression values.

**LncRNA expression during epithelial cell infection.** We assessed the expression of all predicted IncRNAs during the infection model of vaginal epithelial cells interacting with four *Candida* species *C. albicans*, *C. tropicalis*, *C. parapsilosis* and *C. glabrata*. For this, we first calculated the mapped read counts for IncRNAs and known features using FeatureCounts v. 1.6.1402. For each species we performed differential expression analysis using DESeq2 v. 1.26.0 Bioconductor package108, by comparing each time point of infection with 0 h time point control samples using the Wald test. Differential expression calls were performed with the count data of both known features and IncRNAs. LncRNAs with log2 fold change (LFC) > 1.5, and pad (adjusted p value) < 0.01 were considered as differentially expressed (unless specified otherwise). Correction of possible batch effects was done using the RUV v. 1.2.0 Bioconductor package.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Data availability** All publicly available RNA-Seq datasets were retrieved from NCBI SRA database (see Supplementary Data 1 for details of all samples used in this study). Reference genomes, genome annotations, orthology information, GO and PFAM domain tables were obtained from Candida Genome Database. Variant calling data were obtained from Candidamidin database. See “Methods” section for more details. All datasets generated in this study are available at our GitHub page https://github.com/Gabaldonlab/IncRNAs.

**Code availability** The whole data analysis pipeline, i.e., all scripts, datasets, annotations, software versions and auxiliary files to reproduce the results of the study, plots and supplementary materials, are available at our GitHub page https://github.com/Gabaldonlab/IncRNAs. Any updates to the codes and supplementary materials will be available in our GitHub page.

Received: 4 May 2021; Accepted: 30 November 2021; Published online: 16 December 2021.
10. Kutter, C. et al. Rapid turnover of long noncoding RNAs and the evolution of

13. Budak, H., Kaya, S. B. & Cagirici, H. B. Long non-coding RNA in plants in the

14. Choudhary, C. et al. Long non-coding RNAs in insects.

8. Wang, H.-L. V. & Chekanova, J. A. Long Noncoding RNAs in Plants.

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-021-27635-4 | www.nature.com/naturecommunications

References

5. Till, P., Mach, R. L. & Mach-Aigner, A. R. A current view on long noncoding

6. Till, P., Mach, R. L. & Mach-Aigner, A. R. A current view on long noncoding

31. David, L. et al. A high-resolution map of transcription in the yeast genome.

340 (2017).

33. Niederer, R., Hass, E. P. & Zappulla, D. C. Long noncoding RNAs in the

32. Petti, L. E., Martinson, T. & Ecker, J. R. MAT1-1 interaction with Arabidopsis

33. Niederer, R., Hass, E. P. & Zappulla, D. C. Long noncoding RNAs in the

32. Bumgarner, S. L., Dowell, R. D., Grisa

30. Boman, A. P., Pfeifer, C., Burtis, D. M. & St-louis, F. A. The Arabidopsis

31. David, L. et al. A high-resolution map of transcription in the yeast genome.

32. Petti, L. E., Martinson, T. & Ecker, J. R. MAT1-1 interaction with Arabidopsis

33. Niederer, R., Hass, E. P. & Zappulla, D. C. Long noncoding RNAs in the

32. Bumgarner, S. L., Dowell, R. D., Grisa

30. Boman, A. P., Pfeifer, C., Burtis, D. M. & St-louis, F. A. The Arabidopsis

31. David, L. et al. A high-resolution map of transcription in the yeast genome.
61. Hovhannisyan, H. & Gabaldon, T. Transcriptome sequencing approaches to elucidate host–microbe interactions in opportunistic human fungal pathogens. 
       *Fungal Genet. Biol.* 193–235, https://doi.org/10.1007/s10632-017-9713-8 (2018).

62. Sellam, A. et al. Experimental annotation of the human pathogen Candida albicans coding and noncoding transcribed regions using high-resolution tiling arrays. *Genome Biol.* 11, R71 (2010).

63. Denessen, P. D., Schröder, M. S., Higgins, D. G. & Butler, G. Identification of non-coding RNAs in the Candida parapsilosis species group. *PloS One* 11, e0163235 (2016).

64. Linde, J. et al. Defining the transcriptomic landscape of Candida glabrata by RNA-Seq. *Nucleic Acids Res.* 43, 1392–1406 (2015).

65. Pekmezovic, M. et al. Candida pathogen induces protective mitochondria-associated type I interferon signalling and a damage-driven response in vaginal epithelial cells. *Nat. Microbiol.* https://doi.org/10.1038/s41564-021-00875-2 (2021).

66. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295 (2015).

67. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512 (2013).

68. Blevins, W. R. et al. Uncovering de novo gene birth in yeast using deep transcriptomics. *Nat. Commun.* 12, 604 (2021).

69. Hezroni, H. et al. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* 11, 1110–1122 (2015).

70. Kannan, S. et al. Transposable element insertions in long intergenic non-coding RNA genes. *Front Bioeng. Biotechnol.* 3, 71 (2015).

71. Capusta, A. et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PloS Genet.* 9, e1003470 (2013).

72. Nam, J.-W. & Bartel, D. P. Long noncoding RNAs in C. elegans. *Genome Res.* 22, 584–589 (2012).

73. Azlan, A., Obeidat, S. M., Yunus, M. A. & Azzam, G. Systematic identification and characterization of Aedes aegypti long noncoding RNAs (IncRNAs). *Sci. Rep.* 9, 12147 (2019).

74. Farrer, R. A. & Synteny Synteny imaging tool for annotated genome assemblies. *BMC Bioinform.* 18, 507 (2017).

75. Li, L., Stoeckert, C. J. Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189 (2003).

76. Mattei, E., Pietrosemoli, M., Ferré, F. & Helmer-Citterich, M. Web-Beagle: a web server for the alignment of RNA secondary structures. *Nucleic Acids Res.* 43, W493–W497 (2015).

77. Jones, A. N. & Satell, M. Challenges and perspectives for structural biology of IncRNAs—the example of the Xist IncRNA A-repeat. *J. Mol. Cell. Biol.* 11, 845–859 (2019).

78. Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208 (2009).

79. Ross, C. J. et al. Uncovering deeply conserved motif combinations in rapidly evolving noncoding sequences. *Genome Biol.* 22, 29 (2021).

80. Graf, J. & Kretz, M. From structure to function: route to understanding IncRNA mechanism. *Bioessays* 42, e2000277 (2020).

81. Guttman, M. & Rinn, J. L. Modular regulatory principles of large non-coding RNA genes. *Nature* 482, 339–346 (2012).

82. Liao, Q. et al. Large-scale prediction of long non-coding RNA functions in a diverse set of fungal species. *RNA* 2280, 482–485 (2016).

83. Tan, J. Y. et al. cis-Acting complex-trait-associated lincRNA expression correlates with modulation of chromosomal architecture. *Cell Rep.* 18, 2280–2288 (2017).

84. Liu, Y. et al. New signaling pathways govern the host response to C. albicans in various niches. *Genome Res.* 25, 679–689 (2015).

85. Skrzypek, M. S. et al. The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. *Nucleic Acids Res.* 45, D932–D936E (2017).

86. Jarrous, J., Morillon, A. & Pinskaya, M. History, discovery, and classification of IncRNAs. *Adv. Exp. Med. Biol.* 1008, 1–46 (2017).

87. Gao, J. et al. LncRNA DINOR is a virulence factor and global regulator of stress responses in Candida auris. *Nat. Microbiol.* 6, 5056 (2018).

88. Zhao, X. et al. Global identification of Arabidopsis IncRNAs reveals the regulation of MA4F by a natural antisense RNA. *Nat. Commun.* 9, 5056 (2018).

89. Serin, E. A. R., Nijveen, H., Hillhorst, H. W. M. & Ligtwerink, W. Learning from co-expression networks: possibilities and challenges. *Front. Plant Sci.* 7, 444 (2016).

90. Leminen, R., Sugavara, H. & Shumway, M., on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Res.* 39, D19–D21 (2011).
