Am J Transl Res 2024;16(7):3191-3210
www.ajtr.org /ISSN:1943-8141/AJTR0157159

Original Article

Exploring the role of candidalysin in the pathogenicity of *Candida albicans* by gene set enrichment analysis and evolutionary dynamics

Xingchen Zhou[1], Xiaolin Chen[3], Qianglong Pan[3], Shengqi Wang[1], Jing Li[2], Ying Yang[1]

[1]*Bioinformatics Center of AMMS, Beijing Key Laboratory of New Molecular Diagnosis Technologies for Infectious Diseases, Beijing 100850, China;* [2]*School of Life Science and Technology, China Pharmaceutical University, Nanjing 210009, Jiangsu, China;* [3]*Sir Run Run Hospital, Nanjing Medical University, Nanjing 210009, Jiangsu, China*

Received April 12, 2024; Accepted May 20, 2024; Epub July 15, 2024; Published July 30, 2024

Abstract: Aims: To explore the pathogenic mechanisms of *Candida albicans* (*C. albicans*), focusing on its impact on human health, particularly through invasive infections in the gastrointestinal and respiratory tracts. Methods: In this study, we evaluated the demographic and clinical profiles of 7 pneumonia patients. Meanwhile, we used Gene Set Enrichment Analysis (GSEA) and Evolutionary Dynamics method to analyze the role of candidalysin in *C. albicans* pathogenicity. Results: By analyzing genomic data and conducting biomedical text mining, we identified novel mutation sites in the candidalysin coding gene *ECE1-III*, shedding light into the genetic diversity within *C. albicans* strains and their potential implications for antifungal resistance. Our results revealed significant associations between *C. albicans* and respiratory as well as gastrointestinal diseases, emphasizing the fungus's role in the pathogenesis of these diseases. Additionally, we identified a new mutation site in the *C. albicans* strain YF2-5, isolated from patients with pneumonia. This mutation may be associated with its heightened pathogenicity. Conclusion: Our research advances the understanding of *C. albicans* pathogenicity and opens new avenues for developing targeted antifungal therapies. By focusing on the molecular basis of fungal virulence, we aim to contribute to the development of more effective treatment strategies, addressing the challenge of multidrug resistance in invasive fungal infections.

Keywords: Candidalysin, biomedical text mining, evolutionary dynamics, hyphal formation genes, gene set enrichment analysis

Introduction

Invasive *Candida* species, particularly *Candida albicans (C. albicans)*, pose significant threats to human health, manifesting their pathogenicity through various infections, notably within the gastrointestinal and respiratory tracts [1-6]. These infections can lead to severe clinical manifestations, ranging from mucosal infections to bloodstream infections, especially in immunocompromised patients [7-11]. The impact of *Candida* infections on the gastrointestinal and respiratory systems underscores the pressing need for a deeper understanding of their virulence mechanisms and the development of more effective treatment strategies [12-16].

Hyphal formation genes play a pivotal role in the pathogenicity of invasive fungi, serving as essential determinants of virulence in *C. albicans*. These genes facilitate the transition from yeast to hyphal form, a morphological change that is closely associated with the fungus's ability to invade host tissues, evade the immune response, and establish infections [17-28]. The study of hyphal formation-related genes is therefore fundamental to understanding the molecular basis of *Candida* pathogenicity and represents a crucial step towards identifying targets for antifungal therapy [29-31].

Recent advancements in candidalysin research have shed light on its significant role in the virulence of *C. albicans*, particularly in the

https://doi.org/10.62347/IZYM9087

Table 1. Patient information

Patient No.	Strain	Sampling site	Age	Sex	BMI	Disease
1	YF1	Sputum	71	Female	25.8	Pneumonia
2	YF3	Sputum	80	Male	19.5	Pneumonia
3	YF5	Sputum	96	Male	22.85	Pneumonia
4	YF2-1	Sputum	87	Female	-	Pneumonia
5	YF2-3	Sputum	86	Male	-	Pneumonia
6	YF2-5	Sputum	81	Male	-	Pneumonia
7	YF2-11	Sputum	84	Male	15.57	Pneumonia

context of gastrointestinal and respiratory diseases. Candidalysin, a cytolytic peptide toxin secreted during hyphal growth, directly contributes to mucosal damage and inflammation, thereby playing a key role in the pathogenesis of infections [32-39]. Research focusing on candidalysin has highlighted its potential as a biomarker for virulence and a target for therapeutic intervention, offering new perspectives on the management of Candida-induced diseases.

The identification and targeting of virulence factors such as candidalysin offer a promising strategy to combat the emerging issue of multidrug resistance in invasive fungal infections [40-43]. With *C. albicans* increasingly developing resistance to existing antifungal agents, there is a pressing need to explore novel therapeutic targets [44-47]. Focusing on virulence mechanisms rather than fungal survival pathways may help to develop treatments that impair the pathogen's ability to cause disease without directly selecting for resistance, offering a sustainable approach to managing fungal infections.

Our study utilizes a combination of Gene Set Enrichment (GSE) analysis and Evolutionary Dynamics research to explore the genetic basis of candidalysin's role in *C. albicans* pathogenicity. By integrating biomedical text mining with genomic data analysis, this research addresses gaps in our current understanding of *Candida*'s virulence mechanisms. This comprehensive approach not only illuminates the evolutionary dynamics of candidalysin and its associated hyphal formation genes but also identifies potential antifungal candidates, showcasing the study's innovative contributions to the field of microbial pathogenesis.

In conclusion, this research represents a significant step forward in the fight against invasive Candida infections, leveraging cutting-edge analytical techniques to uncover the complex interactions between host and pathogen. By focusing on candidalysin and hyphal formation genes as pivotal elements in *C. albicans*' arsenal of virulence factors, the study provides new insights into the mechanisms driving fungal pathogenicity. This foundational knowledge paves the way for the development of targeted therapies, offering hope for more effective management of invasive fungal diseases in the future.

Materials and methods

Patients

We analyzed the demographic and clinical profiles of 7 pneumonia patients. Patient No. 1, a 71-year-old female, was diagnosed with viral pneumonia, COVID-19, and was post-operative for lung adenocarcinoma and breast cancer. Patient No. 2, an 80-year-old male, suffered from a range of illnesses such as septic shock-induced respiratory failure, COPD, and severe malnutrition, alongside COVID-19. The list included a 96-year-old male. Patient No. 3, with COPD and heart failure, and Patient No. 4, an 87-year-old female, dealing with type I respiratory failure, severe pneumonia, and Alzheimer's disease. Patient No. 5, an 86-year-old male, was diagnosed with COPD, liver and renal dysfunction and suspected of having COVID-19. Patient No. 6, an 81-year-old male, exhibited similar conditions with the addition of coronary heart disease. Lastly, Patient No. 7, an 84-year-old male, was battling type I respiratory failure, suspected COVID-19, with a history of lacunar stroke (**Table 1**). All patients sputum samples were indicative of respiratory tract involvement.

Literature search and screening

A comprehensive search was conducted within the Clarivate Analytics Web of Science Core Collection database, spanning from January 1, 1990 to November 2023. The search string is as follows: 'ALL = (virulence OR VF OR toxin) AND TS = (fungi OR fungus OR mycobiome OR fungal OR fungous) AND LA = (English)'. The

inclusion criteria were restricted to original research articles.

Bibliometric analysis and visualization

Upon data collection, the 'bibliometrix' v4.0.1 package in R software was utilized to perform the analysis [48]. The process began with data cleaning and preprocessing to remove duplicates and correct inconsistencies. Subsequently, the analysis encompassed several dimensions, including publication trends over time, citation analysis, country or region distribution, and the correlation between journals and publications.

Keywords analysis

The identification of gene entries was collected from the NCBI Gene database (https://www.ncbi.nlm.nih.gov/gene), a primary source for molecular biology information on all known and predicted gene sequences. For our study, we focused on 20 gene entries specifically related to hyphal formation, a critical virulence factor in *C. albicans* pathogenesis [49-61]. Fungi and bacteria entries were collected from the NCBI Taxonomy database (https://www.ncbi.nlm.nih.gov/taxonomy). Human disease entries were obtained from the MalaCards human disease database (http://www.malacards.org) [62], which compiles detailed disease-related information from various sources, providing a broad overview of known diseases. For our study, we extracted 22,811 disease entries.

CiteSpace (Citation Space) was used to perform burst detection analysis, which identifies keywords that have shown a sudden increase in usage over time [63]. This analysis can pinpoint emerging trends and shifts in research focus within the domain.

The 'tm' v0.7-8 package in R software was employed to clean and preprocess the text data from publications, which involves removing stopwords, punctuation, and numbers, and performing word stemming to reduce words to their root form [64]. Hypergeometric analysis in R package was performed to identify significantly enriched genes, microbes and human diseases of interest [64, 65].

Genome data collection and genome structure prediction

High-quality genomic data for *C. albicans* were systematically gathered from the NCBI genome database (https://www.ncbi.nlm.nih.gov/genome). This data collection aimed to obtain comprehensive genomic sequences to facilitate detailed analysis. Subsequently, custom scripts were developed and utilized to calculate the genomic size in the genome of each strain. To assess the quality of the genome assemblies, metrics such as N50, GC content percentage and N90 were computed (Table S2).

For the structural prediction of the genomes, Funannotate software (https://github.com/nextgenusfs/funannotate) was employed. The annotation process involved several steps, starting with the 'clean' command to prepare the data by removing contaminants and low-quality sequences. The 'sort' command was then used to order the data in a manner conducive to annotation. The 'mask' command was applied to identify and mask repetitive elements, which can interfere with the annotation process. The 'train' command facilitated the creation of a custom model for gene prediction based on the input data, which is crucial for accurate genome annotation. Finally, the 'predict' command was used to predict the location of genes and their structures within the genome using the default parameters provided by the software.

Pangenome analysis

The identification of homologous protein families within *C. albicans* genomes was conducted using the OrthoFinder2 software [66]. In this context, genes present in all strains of *C. albicans* were classified as core genes. Conversely, genes exclusive to a single strain were designated as strain-specific genes. The remaining genes, not fitting into the core or strain-specific categories, were considered dispensable genes.

To visually represent the distribution and variability of gene counts across different strains, custom scripts were developed to generate rarefaction curves displayed as box-plots.

Multiple sequence alignment phylogenetic tree analysis

Additionally, we employed Mafft v7.508 [67, 68] to align the multiple sequences. A maximum-likelihood phylogenetic tree was then constructed using single-copy orthologs with RAxML v8.2.12 [69], utilizing the PROTCATWAG model and 1000 bootstrap replicates. The multiple sequence alignment was visualized using the 'ggmsa' v1.2.3 [70] package in R. The evolutionary tree was visualized using the 'ggtree' v3.4.4 [71, 72] package in R.

dN/dS value calculation

The dN/dS ratio for each site in the gene was calculated using the SLAC (Single-Likelihood Ancestor Counting) method implemented in HyPhy v2.5.42 [73].

Whole genome sequence

Initially, raw sequencing reads were quality-checked and preprocessed using tools like FastQC and Trimmomatic [32], which allowed us to remove low-quality reads and adapter sequences, ensuring that only high-quality data was used for subsequent analyses.

The preprocessed reads were then aligned to a reference genome using the BWA-MEM [74] algorithm. Post-alignment processing, including sorting and marking duplicates, was performed using Samtools and Picard Tools, respectively, to prepare the data for variant calling.

Variant calling was carried out using the GATK (Genome Analysis Toolkit) Best Practices workflow, which includes steps such as Base Quality Score Recalibration (BQSR) and HaplotypeCaller for identifying single nucleotide polymorphisms (SNPs) and insertions-deletions (indels). This approach is widely recognized for its robustness in detecting genomic variants with high precision.

Fst calculation: Fst is a measure of genetic differentiation among populations, indicating the proportion of genetic variance found between groups relative to the total genetic variance. To compute Fst, we used the VCFtools software, which facilitates the comparison of variant frequency data across populations. For each genomic locus, Fst was calculated using the formula:

$$F_{st} = \frac{\pi_t - (\pi P_1 + \pi P_2)/2}{\pi_t} \quad (1).$$

Where π_t is the nucleotide diversity of the combined populations and π_{p1}, π_{p2} are the nucleotide diversities of each population individually. This approach allows for the identification of genomic regions with high differentiation, potentially indicating selection or local adaptation.

π calculation: Nucleotide diversity, denoted as π, measures the average pairwise differences between individuals within a population. To calculate π, we utilized the PopGenome package in R, an efficient tool for population genetic analyses of large genomic datasets. Nucleotide diversity was calculated using the formula:

$$\pi = \frac{\sum_{i=1}^{n}(x_i)(n - x_i)}{n(n-1)/2} \quad (2).$$

Where n is the sample size, and x_i is the number of copies of the i^{th} allele for the gene. This measure provides insights into the level of genetic variation present within a population, which is crucial for understanding population structure, history, and dynamics.

Both Fst and π were calculated across sliding windows along the genome to generate detailed profiles of genetic differentiation and diversity. This windowed approach allows for the identification of specific regions under selection pressure or with significant genetic variation, facilitating further investigations into their evolutionary significance or association with phenotypic traits.

qPCR analysis

We utilized quantitative PCR (qPCR) to assess the impact of morphological state (hyphal and yeast forms) on the expression levels of three genes: *ECE1*, *HOG1*, and *RHR2* in various strains of *C. albicans*. Initially, strains SC5314, ATCC 18804, ATCC 14053, ATCC 10231, NCYC 4145, and YF2-5 were cultured in Sabouraud dextrose broth overnight and then incubated at 25°C with 200 rpm agitation for 12 hours to promote yeast growth. In contrast, to induce hyphal formation, these six strains were culti-

Table 2. Primer sequences of qPCR genes

Gene	Primer Sequence (5' to 3')
ACT-F	CCAGGTATTGCTGAACGTATGC
ACT-R	GGACCAGATTCGTCGTATTCTTG
ECE1-F	GCCGTCGTCAGATTGCCA
ECE1-R	AACAGTTTCCAGGACGCCA
HOG1-F	GTCTGTGGGTTGTATCTTAG
HOG1-R	TCACTAAATGGGATAGGGTC
RHR2-F	GCCGTACATTTGATGTCATT
RHR2-R	AAAGTACCAGAAGTGACAAC

vated in Spider medium at 37°C with 200 rpm agitation for the same duration.

Following cultivation, cells from both conditions were harvested, washed twice with phosphate-buffered saline (PBS), and subjected to RNA extraction using the hot phenol method. This method involves the addition of hot phenol and chloroform to the cell suspension, which effectively disrupts the cells and allows for the separation of nucleic acids from proteins and lipids by phase separation. After the aqueous phase containing RNA was collected, it was precipitated with isopropanol, washed with ethanol, and resuspended in diethyl pyrocarbonate (DEPC)-treated water to ensure the removal of any potential RNases.

The extracted RNA was then reverse transcribed into cDNA using a reverse transcription kit. This process started with the mixing of RNA, reverse transcriptase, and specific primers in a reaction buffer, followed by incubation at temperatures optimal for enzyme activity to synthesize first-strand cDNA.

Quantitative PCR was subsequently performed using SYBR Green as the fluorescent probe to quantify gene expression. The cycling conditions were optimized for denaturation, annealing, and extension to ensure efficient and specific amplification of the target genes. The housekeeping gene *ACT* was used as an internal control to normalize the expression data and compensate for any variations in cDNA input or amplification efficiency.

The sequences of the gene primers used for the quantitative PCR (qPCR) are provided in **Table 2**.

This methodology allowed us to compare gene expression levels effectively across different fungal states and strains, providing insights into the genetic regulation associated with morphological adaptations in *C. albicans*.

The data results are presented as relative fold change in gene expression between the hyphal and yeast forms of each strain, displayed as mean ± standard deviation. Statistical analysis was performed using the Student's t-test. Statistical significance is denoted as follows: * for $P < 0.05$, ** for $P < 0.01$, and *** for $P < 0.001$.

Results

Bibliometric analysis

Growth model of research trends: In this study, we conducted a comprehensive bibliometric analysis of literature spanning over three decades, from 1992 to 2023, focused on the virulence factors of invasive fungi. To assemble the corpus of relevant research, we meticulously downloaded and curated articles from the Clarivate Analytics Web of Science Core Collection database. Through a rigorous process of filtering and deduplication, we isolated a dataset comprising 22,852 publications for our analysis. This dataset reveals a consistent and noteworthy trend: the volume of research on the virulence factors of invasive fungi has shown a steady annual increase. This increase is particularly evident in the publication volumes for 2022 and 2023, each approaching 1,800 articles (**Figure 1A**). The average annual growth rate of literature over the past five years has remained at 4%, underscoring sustained interest and ongoing research activity within this field (**Figure 1B**).

Despite the thirty years of persistent scholarly attention on invasive fungal virulence factors, the trend in research output also show an exponential growth. The trend is characterized by mathematical expression:

$$f(t) = e^{0.129t - 251} \quad (3).$$

Where t represents time in years. This model is robust, as indicated by a coefficient of determination (R^2) of 0.94 and a highly significant *p*-value of 2.72×10^{-20} (**Figures 1A**, S1). These statistics highlight the accelerating pace of research and the expanding body of knowledge on the virulence factors of invasive fungi.

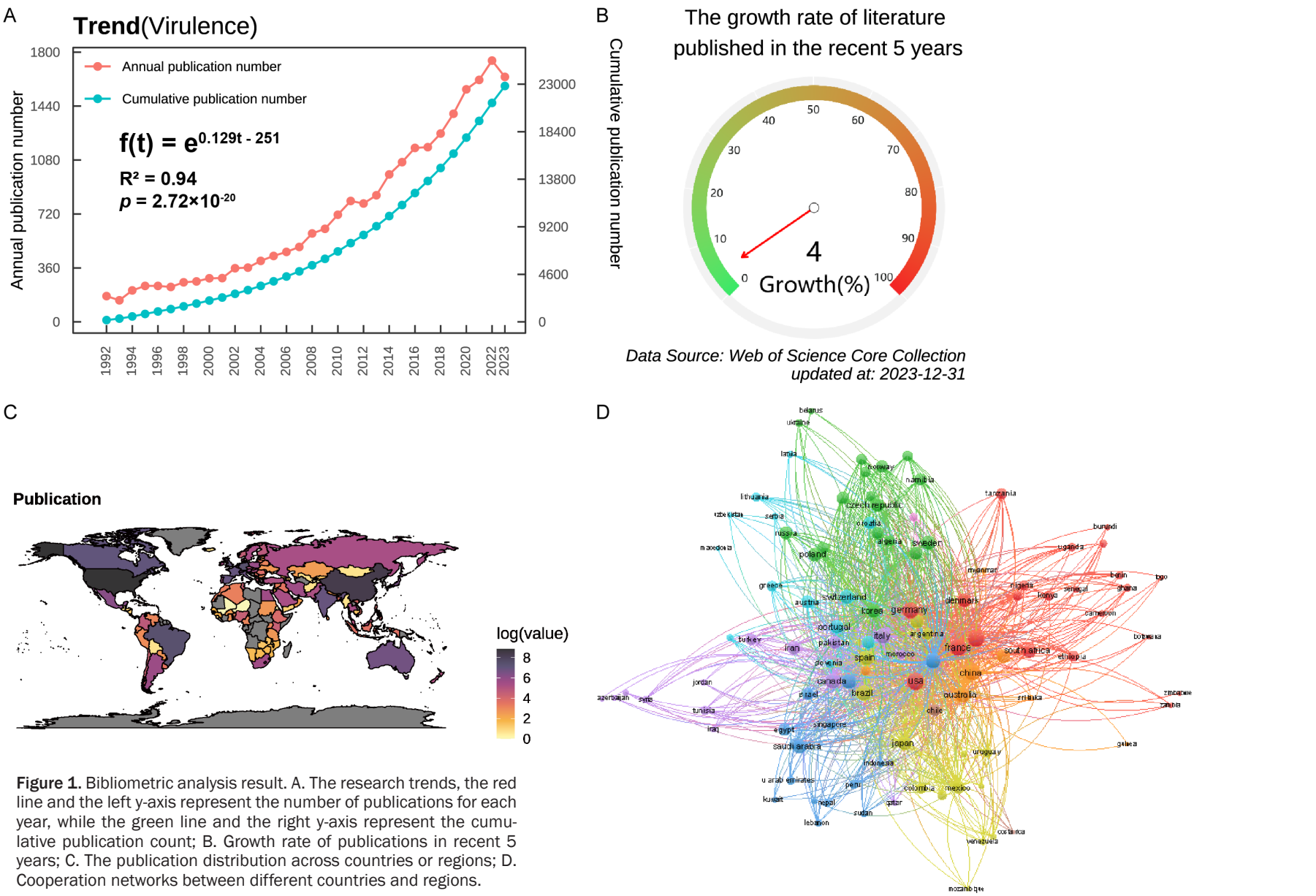


Figure 1. Bibliometric analysis result. A. The research trends, the red line and the left y-axis represent the number of publications for each year, while the green line and the right y-axis represent the cumulative publication count; B. Growth rate of publications in recent 5 years; C. The publication distribution across countries or regions; D. Cooperation networks between different countries and regions.

The distribution of countries or regions and their cooperation networks: The geographical analysis highlights the United States and China as the foremost contributors to research on fungal virulence factors, with the United States leading with 5,231 publications and China following at 79% of this volume with 4,130 publications. The top five countries, including Brazil, Germany, and India, collectively underscore the global engagement in this research area (**Figure 1C**). This distribution reflects not only the scientific focus but also potential collaborations to advance understanding and management of fungal diseases.

Based on the analysis of scientific research collaboration networks among countries, we categorized nations into four distinct clusters. Cluster 1 comprises 34 countries, Cluster 2 contains 37 countries, Cluster 3 includes 18 countries, and Cluster 4 consists of 17 countries. Within Cluster 1, the United States stands out with a betweenness centrality of 483.55, a closeness centrality of 0.00617, and a PageRank centrality of 0.022. These metrics are the highest in the cluster, unequivocally indicating that in the realm of research on virulence factors of invasive fungi, the United States plays a pivotal bridging role in international scientific research collaborations, showcasing the closest degree of cooperative research engagement. Following closely are Germany, Japan, France, and Canada, all demonstrating high levels of collaboration, with China's betweenness centrality slightly lower at 168.45. In Cluster 2, India leads with a betweenness centrality of 191.14. Brazil dominates Cluster 3 with a betweenness centrality of 233.28, while Australia leads Cluster 4 with a betweenness centrality of 535.24, each representing the most collaboratively engaged countries within their respective clusters (**Figure 1D**; Table S1).

Biomedical text mining

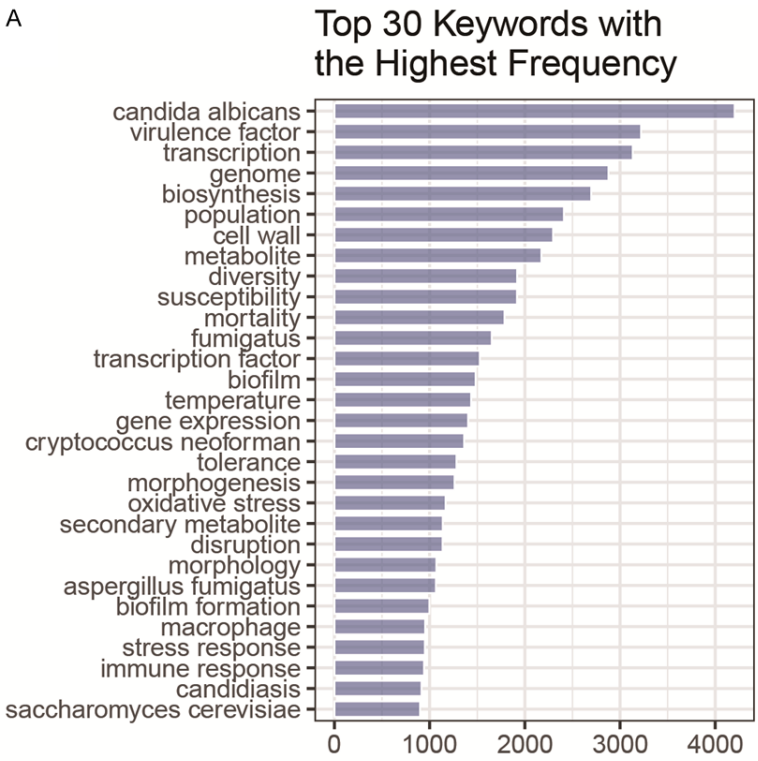
We further engaged in the downloading, text preprocessing, and mining of research papers pertinent to virulence factors of invasive fungi, as well as papers related to fungal research.

The top 30 high-frequency words and burst words: Based on the results of keyword frequency analysis, "*Candida albicans*" undoubtedly emerged as the most frequently occurring keyword. This aligns with the fact that research on *C. albicans* predates other studies in the domain of invasive fungal research, and it is the most detected fungus in hospital-acquired fungal infections. Furthermore, the appearance of keywords such as "genome", "transcription", and "metabolite" indicates that research methodologies for studying virulence factors predominantly focus on genomics, transcription, and metabolomics within the framework of systems biology. Additionally, the frequent occurrence of terms like "cell wall", "biofilm", "temperature", "morphogenesis", and "immune response" suggests that the invasive characteristics and mechanisms of Candida species are associated with these features (**Figure 2A**). For instance, invasive fungi exhibit morphological changes due to alterations in the cell wall and biofilm in response to varying temperatures, thereby facilitating invasion, especially in hosts with compromised immune function.

Analysis of emerging keywords identified "mutant", "signal transduction", and "filamentous growth" as areas receiving substantial attention in early research, including subsequent mentions of "hyphal formation" and "mating type". These findings highlight that hyphal formation and growth are among the most critical characteristics of fungal infections in hosts. In recent years, "*Candida auris*", known as a super-fungus, has also garnered extensive research interest. Additionally, "*Candida glabrata*" and "*Candida parapsilosis*" have been widely studied, indicating that besides *C. albicans*, new Candida species have become prevalent in invasive fungal infections in recent years. This trend is likely linked to environmental changes and adaptive mutations, warranting further investigation. Emerging topics such as "biosynthetic gene cluster" and "extracellular vesicle" have also seen a surge in research focus in recent years as novel biosynthetic pathways and drug delivery mechanisms, respectively (**Figure 2B**).

Gene set enrichment analysis: Disease set enrichment analysis indicates a significant association of *C. albicans* with respiratory and gastrointestinal diseases, with disease ratio of 0.101021 for respiratory diseases and 0.101391 for gastrointestinal diseases, respectively (**Figure 3A**). Studies in larger quantities, significantly enriched within the field of research on virulence factors of invasive fungi,

A



B

Top 30 Keywords with the Strongest Citation Bursts

Keywords	Year	Strength	Begin	End	2000-2023
mutant	2000	31.7069	2000	2007	
signal transduction	2000	35.5114	2000	2008	
filamentous growth	2000	29.8982	2000	2012	
reduced virulence	2000	13.444	2001	2005	
hyphal formation	2000	16.6263	2001	2010	
gene disruption	2000	14.0749	2001	2015	
genetic variation	2000	6.8467	2002	2008	
aspartyl proteinase	2000	7.008	2003	2008	
sexual cycle	2000	8.5659	2004	2011	
siderophore biosynthesis	2000	10.5618	2007	2013	
mating type	2000	6.4021	2008	2015	
avirulence gene	2000	3.8846	2009	2012	
programmed cell death	2000	9.7031	2011	2015	
comparative genomics	2000	4.8434	2015	2018	
azole resistance	2000	6.4464	2016	2021	
invasive candidiasis	2000	3.7051	2017	2018	
stress tolerance	2000	10.171	2017	2020	
fatty acid	2000	8.0107	2017	2020	
host-pathogen interaction	2000	8.0517	2017	2021	
candida glabrata	2000	7.1345	2017	2023	
candida parapsilosis	2000	7.413	2018	2020	
candida spp	2000	7.0345	2018	2021	
fluconazole resistance	2000	6.6025	2018	2023	
inflammation	2000	6.8505	2019	2021	
mutation	2000	5.8068	2019	2021	
candida auri	2000	18.0284	2019	2023	
extracellular vesicle	2000	14.1495	2019	2023	
climate change	2000	9.5953	2019	2023	
biofilm	2000	9.8677	2020	2023	
biosynthetic gene cluster	2000	7.0319	2021	2023	

Figure 2. Biomedical text mining of keywords in publications. A. Top 30 highest frequency keywords. B. Top 30 burst words, the "Strength" represents the citation strength of Burstword, red color represents the burst year from Begin to End.

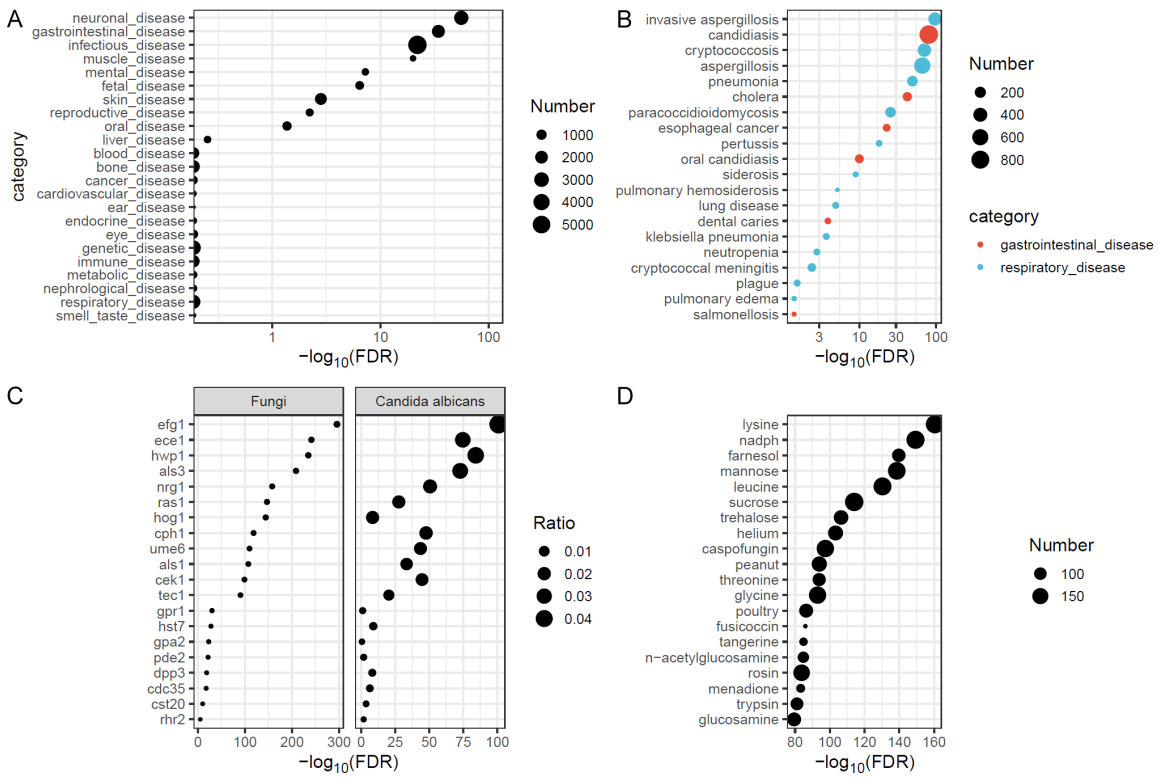


Figure 3. Gene set enrichment analysis. A. Disease set enrichment analysis. B. Enrichment analysis of gastrointestinal disease (red dot) and respiratory disease (blue dot). The y-axis represents disease categories within gastrointestinal or respiratory diseases, while the x-axis represents the *p*-value adjusted by FDR. C. Enrichment analysis of genes related to hyphal formation in fungi or *Candida albicans*, respectively, the dot size represents the ratio of the studied genes to the total number of genes related to fungi or *Candida albicans* research. The x-axis represents the *p*-value adjusted by FDR. D. Drug set enrichment analysis.

also include those related to infectious diseases and neuronal diseases. Conversely, respiratory and gastrointestinal diseases represent two domains within fungal research that are closely associated with specific human body parts, and where research is relatively widespread with more mature methodologies. This suggests a notable prevalence of *C. albicans* in conditions affecting these body systems, reinforcing the importance of understanding its role in pathogenesis and progression of such diseases.

Among the top 20 diseases significantly enriched in studies on gastrointestinal and respiratory illnesses, six pertain to gastrointestinal diseases, while fifteen are related to respiratory conditions. Notably, the diseases with the highest number of studies and significant enrichment are those associated with specific fungal species, such as candidiasis, aspergillosis, cryptococcosis, and others. Following these, pneumonia and other pulmonary diseases are extensively studied. Therefore, it is highly probable that invasive Candida species are among the primary infectious agents in diseases such as pneumonia (**Figure 3B**).

The enrichment analysis on microbial genus and species, including bacteria and fungi, provides insight into the complex interplay between *C. albicans* and other microorganisms (Figure S2). Notably, the analysis of fungal data emphasizes the diverse microbial landscape in which *C. albicans* operates, further highlighting the significance of studying its interactions within the microbiome for potential therapeutic interventions.

A focused enrichment analysis on hyphal formation genes reveals the critical roles of EFG1 and *ECE1* in *C. albicans*. *EFG1*, documented 136 times among 2,765 *C. albicans*-related genes with a significant *p*-value of 6.75×10^{-103}, and ECE1, with 87 records and a *p*-value of 3.87×10^{-76}, emerge as key virulence factors

(**Figure 3C**). Their prominence in the data underscores their potential as targets for therapeutic strategies, given their integral roles in hyphal formation and pathogenicity.

Finally, the enrichment analysis of the drug set revealed that lysine exhibits the highest significance (**Figure 3D**). Lysine can affect the integrity of the cell wall and the metabolic pathways of invasive fungi, thereby influencing their invasiveness. Subsequently, NADPH plays a role in protecting invasive fungi from oxidative stress damage and can promote the synthesis and repair of cell walls and membranes. Farnesol is a crucial substance regulating the morphological transition of invasive fungi, particularly inhibiting the formation of fungal hyphae, thus affecting their invasiveness. Mannose impacts the composition and structure of the cell walls of invasive fungi, influencing the fungus's adhesiveness and invasiveness, as well as its adaptability to various host environments. These substances are pivotal in affecting the pathogenicity of invasive fungi and are key factors in the research of fungal virulence.

In summary, the enrichment analysis conclusively demonstrates the pivotal roles of hyphal formation genes, especially *EFG1* and *ECE1*, in the pathogenicity of *C. albicans*. Their significant association with respiratory and gastrointestinal diseases highlights the impact of *C. albicans* on human health, emphasizing the need for targeted research on these genes. Understanding the mechanisms behind hyphal formation can lead to novel therapeutic approaches to combat infections caused by *C. albicans*.

Population evolutionary analysis of key hyphal formation genes

Candida albicans pan-genome construction: Following an in-depth analysis of keywords extracted from the literature, we curated and examined 45 high-quality genomes of *C. albicans* for comparative genomic studies (Table S2). Our objective was to uncover the evolutionary dynamics of genes implicated in hyphal formation.

Utilizing genomic structure predictions, we charted the proteomic profile of *C. albicans*, uncovering a vast array of protein lengths that span from diminutive peptides of merely 40 amino acids to extensive proteins exceeding 5000 amino acids (Figure S3B). Remarkably, this diversity in protein lengths was consistently observed across various strains of *C. albicans*.

The genomic analysis revealed that *C. albicans* harbors approximately 5,300 coding genes (**Figure 4B**). As our dataset expanded to include additional genomes, the gene family count in *C. albicans* achieved a plateau at 7,242. Within this repertoire, 2,002 families were identified as core gene families, accounting for 27.6% of the total gene pool, whereas 1,168 were delineated as strain-specific genes, representing 16.1%. The balance comprises dispensable genes, a substantial portion of which are prevalent across the majority of *C. albicans* strains. Specifically, 1,453 dispensable genes were detected in 44 strains, representing 20% of the gene families, and 764 genes were present in 43 strains, accounting for 10.5% (**Figures 4A**, **4B**, S3A).

Through presence-absence variation (PAV) analysis, we scrutinized the distribution patterns of 20 genes associated with hyphal formation across different strains of *C. albicans*, aiming to discern potential subspecific differentiation based on these genetic markers. Our analysis indicated that most of the hyphal formation-related genes are almost universally present in all examined *C. albicans* strains, with a minority being absent in 2-3 strains. Specifically, *DPP3*, *HST7*, and *PDE2* were absent in one strain each; *CEK1*, *GPA2*, and *HWP1* were missing in two strains; and *CPH1* was absent in three strains (**Figure 4C**). These findings suggest nuanced genetic variations that may reflect evolutionary adaptations of *C. albicans* in relation to hyphal formation, underscoring the complexity of its genome and the potential for further subspecies differentiation.

Selection pressure analysis of the candidalysin coding gene ECE1-III: Based on the enrichment analysis of genes related to hyphal formation and research over the recent five years on virulence factors associated with hyphal formation (**Figure 3**), transcription regulatory factor *EFG1* and candidalysin coding gene *ECE1* have been identified as key factors in *C. albicans* for hyphal formation and host invasion. *EFG1* acts as a transcriptional regulator within the cAMP-PKA signaling pathway, while *ECE1*, upon cleav-

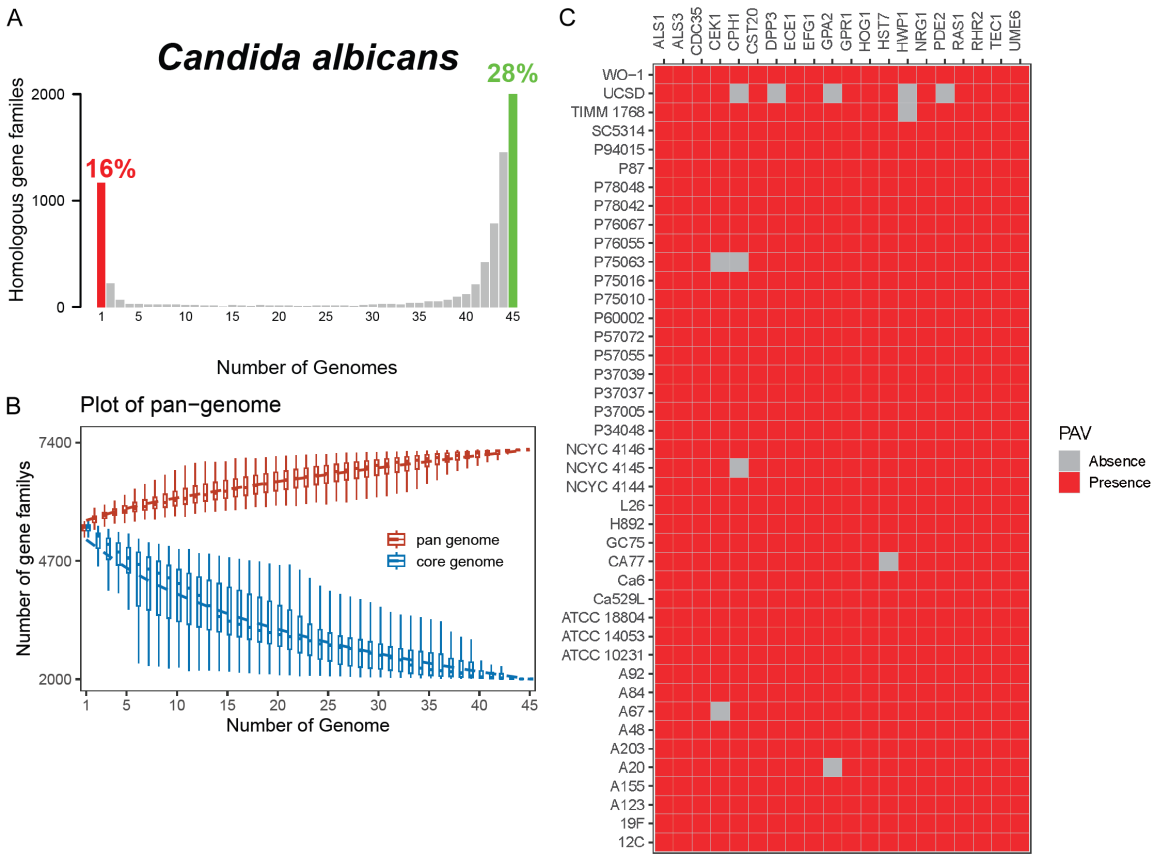


Figure 4. Pangenome analysis of the *Candida albicans* gene family. A. Distribution of core genes (green), dispensable genes (grey) and strain-specific genes (red) in pangenome, the x-axis represents the number of genomes. B. The rarefaction curve of pan-genome (red box) and core-genome (blue box), the x-axis represents the number of genomes, while the y-axis represents the number of gene families. C. Presence/absence variation among the *C. albicans* strains. Red represents presence, while grey represents absence. The top x-axis represents genes, y-axis represents strain.

age and modification by the Kex2/1p enzyme, produces ECE1-III, which ultimately forms candidalysin, posing a threat to human health.

Accordingly, we conducted a comparative analysis of mutation sites specifically on ECE1-III (SIIGIIMGILGNIPQVIQIIMSIVKAFKGNK). Through cluster analysis based on these mutation sites, *C. albicans* strains can be categorized into five groups. The first group (SC5314, A48, A20, A84, P78048, P94015, P37037, 19F, L26, P34048, P57055, P37039, H892, CA77) represents the wild-type sequence. The second group (NCYC 4145, A67, Ca529L, P60002, P78042, A123) primarily exhibits mutations I2F, I3L, G4S, M7T, G8A, I9L, S22G, K28R. The third group (P75010, TIMM 1768, ATCC18804, UCSD) is characterized by the I9L, G11T, K25R mutation. The fourth group (A155, P76067, P76055, ATCC 10231) exhibits mutations G4S, G11S and K25R. The fifth group (WO-1, A92, A203, ATCC14053, GC75, P37005, P57072, 12C, P87, P75063, P75016, Ca6, NCYC 4146, NCYC 4144) is associated with the G4S and K25R mutation. Notably, the second group harbors the highest number of mutations, with eight distinct mutation sites identified (**Figure 5A**).

This detailed mutation site comparison and subsequent cluster analysis of ECE1-III in *C. albicans* not only offer insights into the genetic diversity within this species but also underscore the evolutionary adaptations that may influence pathogenicity. Such findings enrich our understanding of the molecular mechanisms underlying hyphal formation and virulence in *C. albicans*, paving the way for targeted therapeutic strategies against this formidable pathogen.

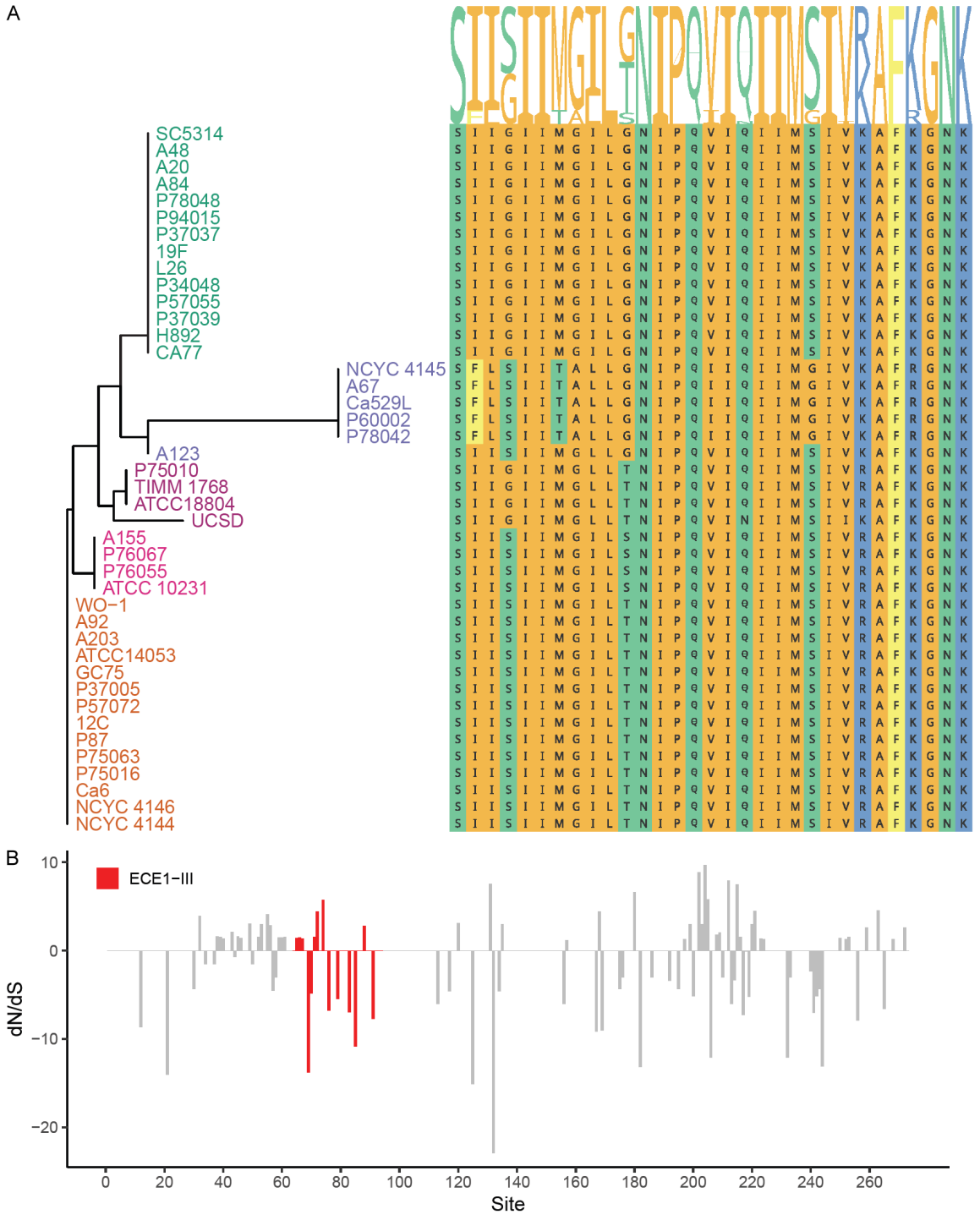


Figure 5. Evolutionary selection pressure analysis. A. On the left is the evolutionary tree of the ECE1-III sequence encoding candidalysin, and on the right is the visualization of motifs and multiple sequence alignment of the sequence, the color of tree label represents different cluster. B. The dN/dS ratio (non-synonymous to synonymous substitution ratio) for each site of ECE1, red color represents ECE1-III (from sites 64 to 94).

Our analysis of the *ECE1* gene using the Single Likelihood Ancestor Counting (SLAC) method revealed a significant pattern of selection pressures when applying a *p*-value threshold of 0.1 [73]. We found no evidence of positive or diversifying selection, as indicated by the absence

of sites with a dN/dS ratio greater than one. In contrast, a substantial number of sites, specifically sixteen, were under negative or purifying selection, suggesting a strong evolutionary constraint to maintain the integrity of the *ECE1* gene (**Figure 5B**; Table S3).

Within the ECE1-III coding region, one particular site, I6, stood out with a notably negative dN/dS ratio of -7.97, and a *p* value of 0.079 for dN/dS being less than one (**Figure 5B**; Table S3). This finding points to a significant purifying selection at this site, emphasizing its potential importance in the gene's function and stability. The negative dN/dS ratio at this site is indicative of strong selective pressure to conserve the amino acid sequence, likely due to its critical role in the protein's function.

Additionally, we identified several sites with dN/dS ratios exceeding one, including I2F, I3L, G8A, I9L, and K25R, with respective values of 1.57, 1.62, 1.68, 3.28, and 1.55 (**Figure 5B**; Table S3). These sites may be subject to positive selection, suggesting that adaptive changes could occur at these positions. However, the presence of these potentially adaptive sites was not sufficient to alter the overall selection pattern of the *ECE1* gene, which is predominantly under purifying selection. The phylogenetic tree constructed from the dN/dS ratio analysis was consistent with the tree derived from mutational data, providing further support for the accuracy of our selection pressure analysis (**Figures 5A**, S4, S5).

Candida albicans isolated from pneumonia patients

Novel mutation sites: We performed whole-genome sequencing on seven *C. albicans* strains isolated from the sputum of pneumonia patients. Additionally, we selected five representative strains from the phylogenetic tree mentioned in subsection 3.3.2.: SC5314, NCYC4145, ATCC 10231, ATCC14053, and ATCC18804 (**Figure 5A**). We extracted their candidalysin coding sequences to further explore their genetic diversity and potential biological significance.

Observations revealed that the candidalysin coding sequences of the YF1 strain showed high similarity with the type culture strain ATCC 14053, indicating great sequence conservation in this particular strain, likely maintaining similar biological functions and pathogenicity. Conversely, the sequence characteristics of YF3 and YF2-1 were closer to another type culture strain, ATCC 18804, suggesting that these strains might share some biological properties or adaptive evolutionary features with ATCC 18804 (**Figure 6A**).

Notably, strains such as YF5, YF2-3, YF2-5, and YF2-11 formed a new cluster of candidalysin coding sequences. The sequence variations within this new cluster, especially mutations at sites I2F, I3L, and G4S, warrant particular attention, as they may be directly linked to the high virulence of the YF2-5 strain (**Figure 6A**).

Selective sweep analysis: In selective sweep analysis, we adopted an integrated approach, focusing on two key population genetics parameters: Fst (fixation index) and π (nucleotide diversity). For precise identification of genes affected by selection, we normalized Fst values via Z-score standardization. Specifically, regions with a Z-score ≥ 1.78 were identified as under selection, employing a statistical threshold to detect genetic markers significantly diverging from the population mean.

Concurrently, we evaluated the π ratio (nucleotide diversity comparison between type culture strains and clinical strains) by computing $\log_2(\pi\_type\_culture/\pi\_clinical)$, establishing two critical thresholds to distinguish selection areas for type culture strains and clinical strains. Selection areas for type culture strains were identified with a ratio ≤ -1.62, whereas areas for clinical strains were determined with a ratio ≥ -0.21. This approach enabled the identification of specific regions impacted by selection, considering genetic diversity.

Using these criteria, we extracted two loci with potential biological significance and annotated their protein functions. Notably, within the selection area of type culture strains, the genes *HOG1* and *RHR2* were found, suggesting potential variations in clinical strains (**Figure 6B**). These mutations might lead to diminished inhibition signals within the HOG signaling pathway, indirectly enhancing the expression of downstream genes. Given the HOG pathway's crucial role in various cellular processes, these mutations could significantly influence the adaptation and evolution of clinical strains.

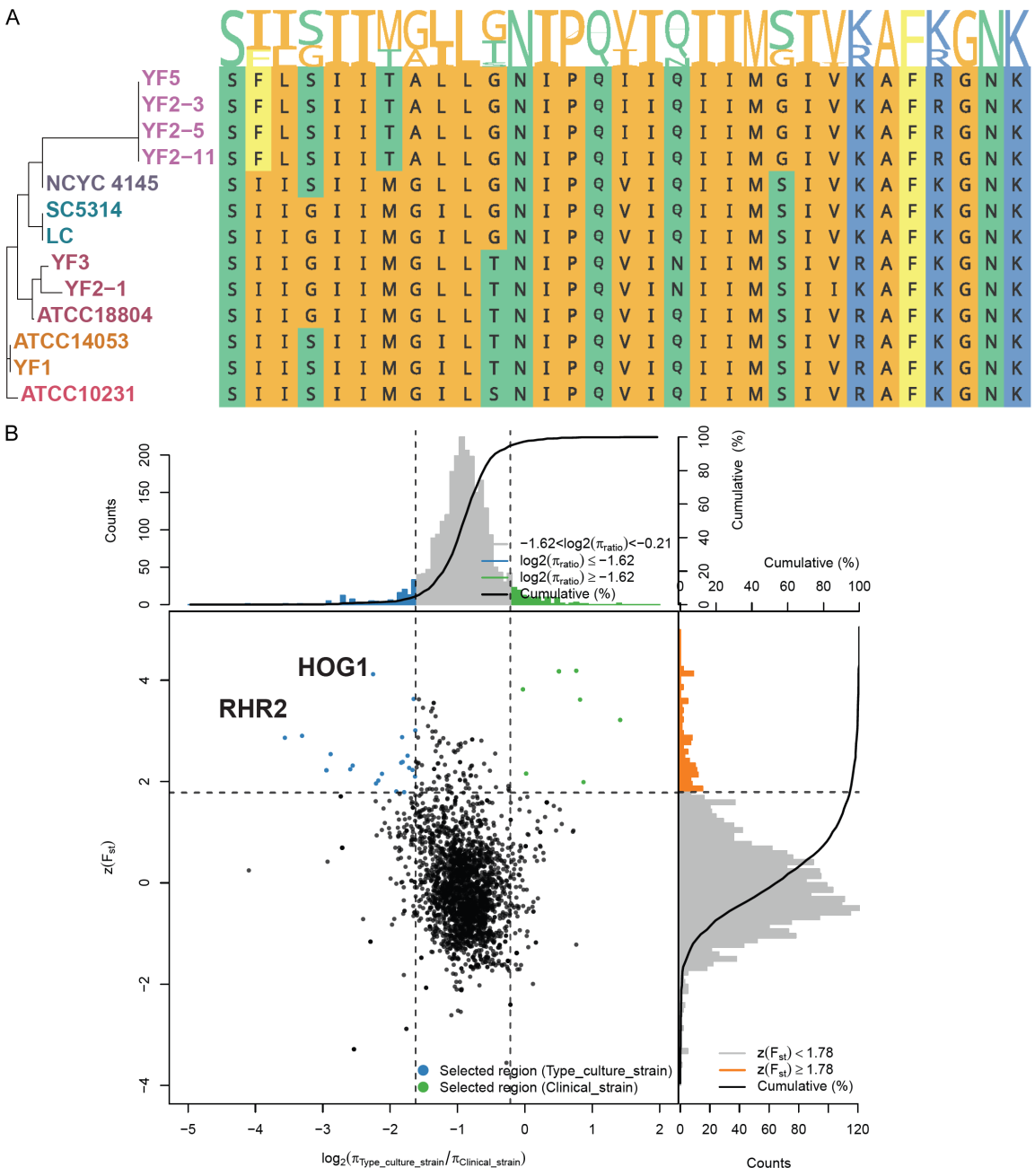


Figure 6. Selective sweep analysis between clinical strain and type culture strain. A. On the left is the evolutionary tree of the ECE1-III sequence encoding candidalysin, and on the right is the visualization of motifs and multiple sequence alignment of the sequence, the color of tree label represents different cluster. All result contains 6 type culture strains and 7 clinical strains. B. Dot plot about the selective sweep analysis result. Blue color points represent selected region of type culture strain, green color points represent selected region of clinical strain, black color points represent another region. X-axis represents $\log_2$(π_type_culture/π_clinical) and y-axis represents z-score of Fst.

This discovery offers valuable clues for further exploration of genetic variations between standard and clinical strains and their effects on pathogenic characteristics.

qPCR analysis

Gene expression levels of *ECE1, HOG1*, and *RHR2* were assessed in representative strains

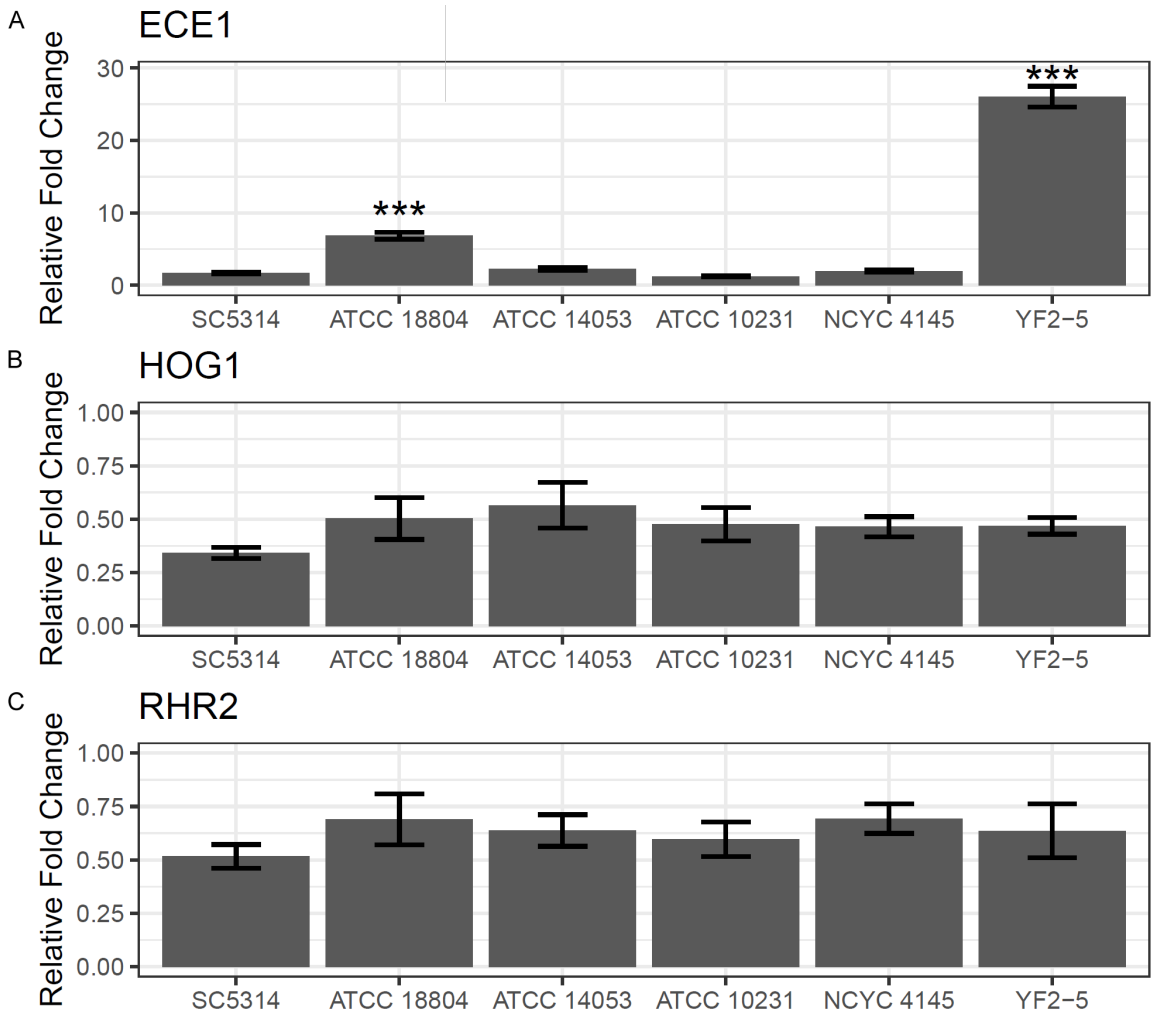


Figure 7. The relative fold changes in the expression levels of the ECE1 (A), HOG1 (B), and RHR2 (C) genes across strains SC5314, ATCC 18804, ATCC 14053, ATCC 10231, NCYC 4145, and YF2-5 of *Candida albicans* (mean ± standard deviation). The size of the error bars represents the standard deviation. Statistical analysis was conducted using the Student's t-test. Statistical significance is denoted by * for $P < 0.05$, ** for $P < 0.01$, and *** for $P < 0.001$.

selected from six different groups of *C. albicans* to investigate the correlation between gene expression and classification based on mutation sites (**Figure 6A**). Additionally, we examined whether the expression level of the ECE1 gene correlates with its clinical origin.

In the analysis of *ECE1* gene expression, the relative fold changes in strains SC5314, ATCC 14053, ATCC 10231, and NCYC 4145 ranged from 1.72 to 1.94, indicating relatively minor variations. However, a significant increase was observed in ATCC 18804, with a fold change of 6.83 ± 0.49 ($P < 0.001$). The clinical isolate YF2-5 showed an even greater significant increase, with a fold change of 26.04 ± 1.45 ($P < 0.001$) (**Figure 7A**). These findings suggest that mutation sites such as I2F and M7T in the ECE1 gene may be associated with higher gene expression and increased pathogenicity.

For the *HOG1* and *RHR2* genes, the relative fold changes of expression levels ranged from 0.34 to 0.56 and 0.51 to 0.69, respectively, with no significant differences observed among the strains (**Figure 7B**, **7C**). The variations between clinical and standard strains may be influenced by factors beyond these two genes.

Discussion

The GSE analysis has significantly advanced our understanding of the virulence factors associated with *C. albicans*, particularly candi-

dalysin. The strong association of *C. albicans* with respiratory and gastrointestinal diseases highlights its role in pathogenesis and progression of these infections. This connection underscores the necessity for targeted research on hyphal formation genes, such as *EFG1* and *ECE1*, which play pivotal roles in the pathogenicity of *C. albicans*. By focusing on these genes, researchers can identify novel therapeutic targets, potentially leading to more effective treatments for *C. albicans* infections.

The population evolutionary analysis of key hyphal formation genes, including the candidalysin coding gene *ECE1*, has provided valuable insights into the genetic diversity within *C. albicans* [75-79]. The identification of novel mutation sites and the classification of *C. albicans* strains based on these mutations suggest a complex evolutionary landscape that may influence pathogenicity and antifungal resistance. Understanding the evolutionary dynamics of candidalysin can inform the development of antifungal agents that are less susceptible to resistance, enhancing the efficacy of treatment strategies against Candida infections [80-82].

The discovery of novel mutation sites in the candidalysin coding sequences from strains isolated from pneumonia patients represents a significant advancement in our understanding of *C. albicans* pathogenicity. These mutations may confer increased virulence or altered antifungal susceptibility, highlighting the importance of genetic profiling in clinical settings. By identifying strains with specific mutations, healthcare providers can tailor treatment strategies to the genetic characteristics of the infecting strain, potentially improving patient outcomes [37, 83, 84].

The geographical analysis of research collaboration networks highlights the importance of international cooperation in advancing our understanding of fungal virulence factors. Collaborative efforts can accelerate the discovery of novel therapeutic targets and antifungal agents, globally pooling resources and expertise. Enhancing collaboration between researchers in this field could lead to breakthroughs in the management and treatment of Candida infections.

We compared the expression levels of the *ECE1*, *HOG1*, and *RHR2* genes among different strains of *C. albicans*. The results indicate that specific mutation sites in strains YF2-5 and ATCC 18804 may lead to increased expression of the ECE1 gene, potentially elevating the pathogenic risk associated with these strains. This novel finding provides a theoretical basis for the detection and targeted therapy of high-risk *C. albicans* strains, as well as offers viable strategies for the prediction and prevention of these pathogens.

However, potential genes *HOG1* and *RHR2* identified through selective sweep analysis did not show any correlation with the clinical origin of the strains, suggesting the involvement of more complex mechanisms. Further research is needed to elucidate these relationships and underlying biological processes.

Future research should continue to explore the genetic basis of *C. albicans* pathogenicity, with a focus on understanding the mechanisms behind hyphal formation and the role of candidalysin in host invasion. Additionally, efforts should be made to identify and characterize novel antifungal compounds through comprehensive drug discovery and development programs. By addressing these key areas, the scientific community can make significant strides in combating Candida infections, improving patient outcomes and reducing the burden of these diseases on healthcare systems worldwide.

Conclusions

This study has significantly advanced our understanding of the pathogenicity of *C. albicans*, particularly focusing on candidalysin's role in respiratory and gastrointestinal diseases through Gene Set Enrichment analysis and Evolutionary Dynamics research. By identifying novel mutation sites and uncovering strain-specific genetic diversity, our findings underscore the importance of targeting hyphal formation genes for novel therapeutic developments. The collaborative nature of this research highlights the global effort required to combat fungal infections effectively. Future studies should aim to translate these genetic and evolutionary insights into practical applications, enhancing the diagnosis, treatment, and management of candidiasis, thereby contributing to improved public health outcomes.

Acknowledgements

This research was funded by the Ministry of Science and Technology of the People's Republic of China, grant number 2018ZX10-101003, the Consulting Research Project of the Chinese Academy of Engineering, grant number 2020-XY-61-03 and the National Natural Science Foundation of China, grant number 32170062.

Disclosure of conflict of interest

None.

Address correspondence to: Ying Yang, Bioinformatics Center of AMMS, Beijing Key Laboratory of New Molecular Diagnosis Technologies for Infectious Diseases, Beijing 100850, China. Tel: +86-010-68170344; E-mail: y_ying_77@163.com

References

[1] Li XV, Leonardi I and Iliev ID. Gut mycobiota in immunity and inflammatory disease. Immunity 2019; 50: 1365-1379.

[2] Bekkers M, Stojkovic B and Kaiko GE. Mining the microbiome and microbiota-derived molecules in inflammatory bowel disease. Int J Mol Sci 2021; 22: 11243.

[3] Liu S, Zhao W, Lan P and Mou X. The microbiome in inflammatory bowel diseases: from pathogenesis to therapy. Protein Cell 2021; 12: 331-345.

[4] Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, Andrews E, Ajami NJ, Bonham KS, Brislawn CJ, Casero D, Courtney H, Gonzalez A, Graeber TG, Hall AB, Lake K, Landers CJ, Mallick H, Plichta DR, Prasad M, Rahnavard G, Sauk J, Shungin D, Vázquez-Baeza Y, White RA 3rd; IBDMDB Investigators; Braun J, Denson LA, Jansson JK, Knight R, Kugathasan S, McGovern DPB, Petrosino JF, Stappenbeck TS, Winter HS, Clish CB, Franzosa EA, Vlamakis H, Xavier RJ and Huttenhower C. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature 2019; 569: 655-662.

[5] Seelbinder B, Lohinai Z, Vazquez-Uribe R, Brunke S, Chen X, Mirhakkak M, Lopez-Escalera S, Dome B, Megyesfalvi Z, Berta J, Galffy G, Dulka E, Wellejus A, Weiss GJ, Bauer M, Hube B, Sommer MOA and Panagiotou G. Candida expansion in the gut of lung cancer patients associates with an ecological signature that supports growth under dysbiotic conditions. Nat Commun 2023; 14: 2673.

[6] Engku Nasrullah Satiman EAF, Ahmad H, Ramzi AB, Abdul Wahab R, Kaderi MA, Wan Harun WHA, Dashper S, McCullough M and Arzmi MH. The role of Candida albicans candidalysin ECE1 gene in oral carcinogenesis. J Oral Pathol Med 2020; 49: 835-841.

[7] Gintjee TJ, Donnelley MA and Thompson GR 3rd. Aspiring antifungals: review of current antifungal pipeline developments. J Fungi (Basel) 2020; 6: 28.

[8] Zhang L, Zhan H, Xu W, Yan S and Ng SC. The role of gut mycobiome in health and diseases. Therap Adv Gastroenterol 2021; 14: 17562848211047130.

[9] Yu D, Xu C, Tu H, Ye A and Wu L. miR-384-5p regulates inflammation in Candida albicans-induced acute lung injury by downregulating PGC1β and enhancing the activation of Candida albicans-triggered signaling pathways. Sci Prog 2021; 104: 368504211014361.

[10] Paramythiotou E, Frantzeskaki F, Flevari A, Armaganidis A and Dimopoulos G. Invasive Fungal Infections in the ICU: how to approach, how to treat. Molecules 2014; 19: 1085-1119.

[11] Shoham S and Marwaha S. Invasive fungal infections in the ICU. J Intensive Care Med 2010; 25: 78-92.

[12] Alonso-Roman R, Last A, Mirhakkak MH, Sprague JL, Möller L, Großmann P, Graf K, Gratz R, Mogavero S, Vylkova S, Panagiotou G, Schäuble S, Hube B and Gresnigt MS. Lactobacillus rhamnosus colonisation antagonizes Candida albicans by forcing metabolic adaptations that compromise pathogenicity. Nat Commun 2022; 13: 3192.

[13] Yan Z, Chen B, Yang Y, Yi X, Wei M, Ecklu-Mensah G, Buschmann MM, Liu H, Gao J, Liang W, Liu X, Yang J, Ma W, Liang Z, Wang F, Chen D, Wang L, Shi W, Stampfli MR, Li P, Gong S, Chen X, Shu W, El-Omar EM, Gilbert JA, Blaser MJ, Zhou H, Chen R and Wang Z. Multi-omics analyses of airway host-microbe interactions in chronic obstructive pulmonary disease identify potential therapeutic interventions. Nat Microbiol 2022; 7: 1361-1375.

[14] Hu Y, Cheng M, Liu B, Dong J, Sun L, Yang J, Yang F, Chen X and Jin Q. Metagenomic analysis of the lung microbiome in pulmonary tuberculosis - a pilot study. Emerg Microbes Infect 2020; 9: 1444-1452.

[15] Tsay JJ, Wu BG, Badri MH, Clemente JC, Shen N, Meyn P, Li Y, Yie TA, Lhakhang T, Olsen E, Murthy V, Michaud G, Sulaiman I, Tsirigos A, Heguy A, Pass H, Weiden MD, Rom WN, Sterman DH, Bonneau R, Blaser MJ and Segal LN. Airway microbiota is associated with upregulation of the PI3K pathway in lung cancer. Am J Respir Crit Care Med 2018; 198: 1188-1198.

[16] Ren L, Wang Y, Zhong J, Li X, Xiao Y, Li J, Yang J, Fan G, Guo L, Shen Z, Kang L, Shi L, Li Q, Li J, Di L, Li H, Wang C, Wang Y, Wang X, Zou X,

Rao J, Zhang L, Wang J, Huang Y, Cao B, Wang J and Li M. Dynamics of the upper respiratory tract microbiota and its association with mortality in COVID-19. Am J Respir Crit Care Med 2021; 204: 1379-1390.

[17] Fazly A, Jain C, Dehner AC, Issi L, Lilly EA, Ali A, Cao H, Fidel PL Jr, Rao RP and Kaufman PD. Chemical screening identifies filastatin, a small molecule inhibitor of Candida albicans adhesion, morphogenesis, and pathogenesis. Proc Natl Acad Sci U S A 2013; 110: 13594-13599.

[18] Vargas-Blanco D, Lynn A, Rosch J, Noreldin R, Salerni A, Lambert C and Rao RP. A pre-therapeutic coating for medical devices that prevents the attachment of Candida albicans. Ann Clin Microbiol Antimicrob 2017; 16: 41.

[19] Mukaremera L, Lee KK, Mora-Montes HM and Gow NAR. Candida albicans yeast, pseudohyphal, and hyphal morphogenesis differentially affects immune recognition. Front Immunol 2017; 8: 629.

[20] Liu H, Köhler J and Fink GR. Suppression of hyphal formation in Candida albicans by mutation of a STE12 homolog. Science 1994; 266: 1723-1726.

[21] Su C, Yu J and Lu Y. Hyphal development in Candida albicans from different cell states. Curr Genet 2018; 64: 1239-1243.

[22] Vediyappan G, Dumontet V, Pelissier F and d'Enfert C. Gymnemic acids inhibit hyphal growth and virulence in Candida albicans. PLoS One 2013; 8: e74189.

[23] Arkowitz RA and Bassilana M. Recent advances in understanding Candida albicans hyphal growth. F1000Res 2019; 8: F1000 Faculty Rev-700.

[24] Ponde NO, Lortal L, Ramage G, Naglik JR and Richardson JP. Candida albicans biofilms and polymicrobial interactions. Crit Rev Microbiol 2021; 47: 91-111.

[25] Lopes JP and Lionakis MS. Pathogenesis and virulence of Candida albicans. Virulence 2022; 13: 89-121.

[26] Mayer FL, Wilson D and Hube B. Candida albicans pathogenicity mechanisms. Virulence 2013; 4: 119-128.

[27] Konno N, Ishii M, Nagai A, Watanabe T, Ogasawara A, Mikami T and Matsumoto T. Mechanism of Candida albicans transformation in response to changes of pH. Biol Pharm Bull 2006; 29: 923-926.

[28] Chen H, Zhou X, Ren B and Cheng L. The regulation of hyphae growth in Candida albicans. Virulence 2020; 11: 337-348.

[29] Liu J, Taft DH, Maldonado-Gomez MX, Johnson D, Treiber ML, Lemay DG, DePeters EJ and Mills DA. The fecal resistome of dairy cattle is associated with diet during nursing. Nat Commun 2019; 10: 4406.

[30] Shuai M, Zhang G, Zeng FF, Fu Y, Liang X, Yuan L, Xu F, Gou W, Miao Z, Jiang Z, Wang JT, Zhuo LB, Chen YM, Ju F and Zheng JS. Human gut antibiotic resistome and progression of diabetes. Adv Sci (Weinh) 2022; 9: e2104965.

[31] Heianza Y, Zheng Y, Ma W, Rimm EB, Albert CM, Hu FB, Rexrode KM, Manson JE and Qi L. Duration and life-stage of antibiotic use and risk of cardiovascular events in women. Eur Heart J 2019; 40: 3838-3845.

[32] Bolger AM, Lohse M and Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics 2014; 30: 2114-2120.

[33] Jolley KA, Bray JE and Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.Org website and their applications. Wellcome Open Res 2018; 3: 124.

[34] Cameron BH and Luong AU. New developments in allergic fungal rhinosinusitis pathophysiology and treatment. Am J Rhinol Allergy 2023; 37: 214-220.

[35] Nishikawa Y, Tomotake Y, Kawano H, Naruishi K, Kido JI, Hiroshima Y, Murakami A, Ichikawa T and Yumoto H. Effects of candidalysin derived from Candida albicans on the expression of pro-inflammatory mediators in human gingival fibroblasts. Int J Mol Sci 2023; 24: 3256.

[36] Chu H, Duan Y, Lang S, Jiang L, Wang Y, Llorente C, Liu J, Mogavero S, Bosques-Padilla F, Abraldes JG, Vargas V, Tu XM, Yang L, Hou X, Hube B, Stärkel P and Schnabl B. The Candida albicans exotoxin candidalysin promotes alcohol-associated liver disease. J Hepatol 2020; 72: 391-400.

[37] Tso GHW, Reales-Calderon JA, Tan ASM, Sem X, Le GTT, Tan TG, Lai GC, Srinivasan KG, Yurieva M, Liao W, Poidinger M, Zolezzi F, Rancati G and Pavelka N. Experimental evolution of a fungal pathogen into a gut symbiont. Science 2018; 362: 589-595.

[38] Ho J, Yang X, Nikou SA, Kichik N, Donkin A, Ponde NO, Richardson JP, Gratacap RL, Archambault LS, Zwirner CP, Murciano C, Henley-Smith R, Thavaraj S, Tynan CJ, Gaffen SL, Hube B, Wheeler RT, Moyes DL and Naglik JR. Candidalysin activates innate epithelial immune responses via epidermal growth factor receptor. Nat Commun 2019; 10: 2297.

[39] Li XV, Leonardi I, Putzel GG, Semon A, Fiers WD, Kusakabe T, Lin WY, Gao IH, Doron I, Gutierrez-Guerrero A, DeCelie MB, Carriche GM, Mesko M, Yang C, Naglik JR, Hube B, Scherl EJ and Iliev ID. Immune regulation by fungal strain diversity in inflammatory bowel disease. Nature 2022; 603: 672-678.

[40] Ali S, Khan FI, Mohammad T, Lan D, Hassan MI and Wang Y. Identification and evaluation of inhibitors of lipase from malassezia restricta

using virtual high-throughput screening and molecular dynamics studies. Int J Mol Sci 2019; 20: 884.
[41] Founta K, Dafou D, Kanata E, Sklaviadis T, Zanos TP, Gounaris A and Xanthopoulos K. Gene targeting in amyotrophic lateral sclerosis using causality-based feature selection and machine learning. Mol Med 2023; 29: 12.
[42] Pun FW, Liu BHM, Long X, Leung HW, Leung GHD, Mewborne QT, Gao J, Shneyderman A, Ozerov IV, Wang J, Ren F, Aliper A, Bischof E, Izumchenko E, Guan X, Zhang K, Lu B, Rothstein JD, Cudkowicz ME and Zhavoronkov A. Identification of therapeutic targets for amyotrophic lateral sclerosis using PandaOmics - An AI-enabled biological target discovery platform. Front Aging Neurosci 2022; 14: 914017.
[43] Eisenstein M. Machine learning powers biobank-driven drug discovery. Nat Biotechnol 2022; 40: 1303-1305.
[44] Ho LT and Goethals PLM. Opportunities and challenges for the sustainability of lakes and reservoirs in relation to the sustainable development goals (SDGs). Water 2019; 11: 23-28.
[45] Dardonville M, Urruty N, Bockstaller C and Therond O. Influence of diversity and intensification level on vulnerability, resilience and robustness of agricultural systems. Agricultural Systems 2020; 2020: 184.
[46] Waugh N, Royle P, Craigie I, Ho V, Pandit L, Ewings P, Adler A, Helms P and Sheldon C. Screening for cystic fibrosis-related diabetes: a systematic review. Health Technol Assess 2012; 16: iii-iv, 1-179.
[47] Demmer S, Kirkman K and Tedder M. What evidence is available on the drivers of grassland ecosystem stability across a range of outcome measurements: a systematic map protocol. Environmental Evidence 2018; 7.
[48] Aria M and Cuccurullo C. Bibliometrix: an R-tool for comprehensive science mapping analysis. J Informetr 2017; 11: 959-975.
[49] Sudbery PE. Growth of Candida albicans hyphae. Nat Rev Microbiol 2011; 9: 737-748.
[50] Li M, Martin SJ, Bruno VM, Mitchell AP and Davis DA. Candida albicans Rim13p, a protease required for Rim101p processing at acidic and alkaline pHs. Eukaryot Cell 2004; 3: 741-751.
[51] Alonso GC, Klein MI, Jordão CC, Carmello JC and Pavarina AC. Gene expression of Candida albicans strains isolates from patients with denture stomatitis submitted to treatments with photodynamic therapy and nystatin. Photodiagnosis Photodyn Ther 2021; 35: 102292.
[52] Hnisz D, Bardet AF, Nobile CJ, Petryshyn A, Glaser W, Schöck U, Stark A and Kuchler K. A histone deacetylase adjusts transcription kinetics at coding sequences during Candida albicans morphogenesis. PLoS Genet 2012; 8: e1003118.
[53] Alonso GC, Pavarina AC, Sousa TV and Klein MI. A quest to find good primers for gene expression analysis of Candida albicans from clinical samples. J Microbiol Methods 2018; 147: 1-13.
[54] Witchley JN, Penumetcha P, Abon NV, Woolford CA, Mitchell AP and Noble SM. Candida albicans morphogenesis programs control the balance between gut commensalism and invasive infection. Cell Host Microbe 2019; 25: 432-443, e6.
[55] Witchley JN, Basso P, Brimacombe CA, Abon NV and Noble SM. Recording of DNA-binding events reveals the importance of a repurposed Candida albicans regulatory network for gut commensalism. Cell Host Microbe 2021; 29: 1002-1013, e9.
[56] Khan F, Bamunuarachchi NI, Tabassum N, Jo DM, Khan MM and Kim YM. Suppression of hyphal formation and virulence of Candida albicans by natural and synthetic compounds. Biofouling 2021; 37: 626-655.
[57] Morschhäuser J. Regulation of white-opaque switching in Candida albicans. Med Microbiol Immunol 2010; 199: 165-172.
[58] Morrow B, Srikantha T, Anderson J and Soll DR. Coordinate regulation of two opaque-phase-specific genes during white-opaque switching in Candida albicans. Infect Immun 1993; 61: 1823-1828.
[59] Lockhart SR, Zhao R, Daniels KJ and Soll DR. Alpha-pheromone-induced "Shmooing" and gene regulation require white-opaque switching during Candida albicans mating. Eukaryot Cell 2003; 2: 847-855.
[60] Zhang A, Liu Z and Myers LC. Differential regulation of white-opaque switching by individual subunits of Candida albicans mediator. Eukaryot Cell 2013; 12: 1293-1304.
[61] Qasim MN, Valle Arevalo A, Nobile CJ and Hernday AD. The roles of chromatin accessibility in regulating the Candida albicans white-opaque phenotypic switch. J Fungi (Basel) 2021; 7: 37.
[62] Rappaport N, Twik M, Plaschkes I, Nudel R, Iny Stein T, Levitt J, Gershoni M, Morrey CP, Safran M and Lancet D. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Res 2017; 45: D877-D887.
[63] Chen C. CiteSpace II: detecting and visualizing emerging trends and transient patterns in scientific literature. J Am Soc Inf Sci Technol 2006; 57: 359-377.
[64] Ren J, Lv Y, Wu L, Chen S, Lei C, Yang D, Li F, Liu C and Zheng Y. Key ferroptosis-related genes in

abdominal aortic aneurysm formation and rupture as determined by combining bioinformatics techniques. Front Cardiovasc Med 2022; 9: 875434.
[65] Grassi M and Tarantino B. SEMgsa: topology-based pathway enrichment analysis with structural equation models. BMC Bioinformatics 2022; 23: 344.
[66] Alhajj MN, Halboub E, Yacob N, Al-Maweri SA, Ahmad SF, Celebić A, Al-Mekhlafi HM and Salleh NM. Adhesion of Candida albicans to digital versus conventional acrylic resins: a systematic review and meta-analysis. BMC Oral Health 2024; 24: 303.
[67] Katoh K and Standley DM. MAFFT: iterative refinement and additional methods. Methods Mol Biol 2014; 1079: 131-146.
[68] Katoh K and Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 2013; 30: 772-780.
[69] Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 2014; 30: 1312-1313.
[70] Zhou L, Feng T, Xu S, Gao F, Lam TT, Wang Q, Wu T, Huang H, Zhan L, Li L, Guan Y, Dai Z and Yu G. Ggmsa: a visual exploration tool for multiple sequence alignment and associated data. Brief Bioinform 2022; 23: bbac222.
[71] Yu G, Smith DK, Zhu H, Guan Y and Lam TT. Ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol 2017; 8: 28-36.
[72] Yu G, Lam TT, Zhu H and Guan Y. Two methods for mapping and visualizing associated data on phylogeny using ggtree. Mol Biol Evol 2018; 35: 3041-3043.
[73] Kosakovsky Pond SL and Frost SD. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 2005; 22: 1208-1222.
[74] Mohan K, Kontz B, Okello P, Allen TW, Bergstrom GC, Bissonnette K, Bonkowski J, Bradley CA, Buck J, Chilvers MI, Dorrance A, Giesler L, Kelly H, Koehler A, Lopez-Nicora HD, Mangel D, Markell SG, Mueller D, Price PP, Rojas A, Shires M, Smith D, Spurlock T, Webster RK, Wise K, Yabwalo D and Mathew FM. Variation in Isolate Virulence and Accession Resistance Associated with Diaporthe Aspalathi D. Caulivora and D. Longicolla in Soybean. Plant Health Progress 2023; 24: 482-487.
[75] Javvaji CK, Reddy H, Vagha JD, Taksande A, Kommareddy A and Reddy NS. Immersive innovations: exploring the diverse applications of virtual reality (VR) in healthcare. Cureus 2024; 16: e56137.
[76] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C and Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018; 46: D1074-D1082.
[77] Probst D and Reymond JL. Exploring DrugBank in virtual reality chemical space. J Chem Inf Model 2018; 58: 1731-1735.
[78] Martino M, Salvadori A, Lazzari F, Paoloni L, Nandi S, Mancini G, Barone V and Rampino S. Chemical promenades: exploring potential-energy surfaces with immersive virtual reality. J Comput Chem 2020; 41: 1310-1323.
[79] Xiong G, Wu Z, Yi J, Fu L, Yang Z, Hsieh C, Yin M, Zeng X, Wu C, Lu A, Chen X, Hou T and Cao D. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. Nucleic Acids Res 2021; 49: W5-W14.
[80] Xie P. Study of international anticancer research trends via co-word and document co-citation visualization analysis. Scientometrics 2015; 105: 611-622.
[81] Santillan-Fernandez A, Salinas-Moreno Y, Valdez-Lazalde JR and Pereira-Lorenzo S. Spatial-temporal evolution of scientific production about genetically modified maize. Agriculture-Basel 2021; 11: 246.
[82] Zwama G, Diaconu K, Voce AS, O'May F, Grant AD and Kielmann K. Health system influences on the implementation of tuberculosis infection prevention and control at health facilities in low-income and middle-income countries: a scoping review. BMJ Glob Health 2021; 6: e004735.
[83] Odds FC. Molecular phylogenetics and epidemiology of Candida albicans. Future Microbiol 2010; 5: 67-79.
[84] Ropars J, Maufrais C, Diogo D, Marcet-Houben M, Perin A, Sertour N, Mosca K, Permal E, Laval G, Bouchier C, Ma L, Schwartz K, Voelz K, May RC, Poulain J, Battail C, Wincker P, Borman AM, Chowdhary A, Fan S, Kim SH, Le Pape P, Romeo O, Shin JH, Gabaldon T, Sherlock G, Bougnoux ME and d'Enfert C. Gene flow contributes to diversification of the major fungal pathogen Candida albicans. Nat Commun 2018; 9: 2253.

Table S1. Countries or regions cooperation network information

vertex	cluster	betweenness_centrality	closeness_centrality	pagerank_centrality
usa	1	483.5485294	0.00617284	0.022008045
china	1	168.4529165	0.005681818	0.017854603
germany	1	314.6151959	0.005952381	0.017102044
united kingdom	1	307.226197	0.005882353	0.020309644
france	1	245.6511225	0.00591716	0.019774129
canada	1	210.3875859	0.005681818	0.016147932
netherlands	1	181.5595982	0.005747126	0.013909039
turkey	1	57.66456914	0.004950495	0.009854406
kenya	1	7.662848367	0.00456621	0.012326546
tunisia	1	1.781245988	0.004464286	0.008567291
nigeria	1	6.288499245	0.00462963	0.008045976
ethiopia	1	39.36155309	0.004739336	0.010086518
philippines	1	6.407363826	0.004484305	0.009686962
iraq	1	0.345742964	0.004149378	0.005008426
bangladesh	1	14.57249608	0.004587156	0.007707645
morocco	1	2.974748426	0.004347826	0.008851604
cameroon	1	0.875716949	0.003984064	0.00656864
benin	1	3.889742152	0.0041841	0.007102413
uganda	1	2.805647612	0.004201681	0.008120367
syria	1	11.12310902	0.004405286	0.007704149
kazakhstan	1	6.406859388	0.003875969	0.005281761
ghana	1	4.568717006	0.004329004	0.007178613
sri lanka	1	0.185235499	0.003546099	0.00411691
burkina faso	1	0.580460637	0.004081633	0.00796184
costa rica	1	0	0.003496503	0.003602537
guinea	1	0	0.003861004	0.002047872
senegal	1	0.599932646	0.003623188	0.006833245
kyrgyzstan	1	0	0.000829876	0.004442561
togo	1	0	0.000841751	0.004600181
botswana	1	0	0.003861004	0.003635505
zambia	1	5.645112377	0.004098361	0.003253269
azerbaijan	1	2.960195587	0.003533569	0.005146907
burundi	1	0.111742424	0.003448276	0.005150999
myanmar	1	0.4786028	0.003546099	0.0042118
india	2	191.1440068	0.005235602	0.015223198
spain	2	63.13687576	0.005076142	0.015701151
italy	2	190.6898854	0.005235602	0.016966445
korea	2	47.70859679	0.004878049	0.014068518
switzerland	2	153.6448613	0.005102041	0.014898783
iran	2	40.63176969	0.004878049	0.014437722
poland	2	41.89508899	0.004807692	0.014972832
austria	2	52.03385556	0.004901961	0.011752965
thailand	2	120.2597405	0.004975124	0.014144455
belgium	2	93.03557257	0.005025126	0.015592467
hungary	2	4.707478836	0.004587156	0.014058302
south africa	2	115.0554045	0.005076142	0.015211829
russia	2	18.01755089	0.004608295	0.015182702

pakistan	2	79.02008834	0.004784689	0.013494203
denmark	2	40.21950335	0.004854369	0.015635254
sweden	2	61.45712967	0.004950495	0.015751874
czech republic	2	32.78713897	0.004784689	0.01578012
new zealand	2	8.132901242	0.004761905	0.013948421
finland	2	7.995464394	0.004651163	0.01342831
chile	2	40.90723457	0.004716981	0.010863044
norway	2	3.639748302	0.004545455	0.013713163
slovakia	2	0.580141561	0.004444444	0.012616872
croatia	2	22.13516557	0.004587156	0.013292773
vietnam	2	19.09735333	0.004545455	0.013685861
lithuania	2	7.153153335	0.004366812	0.007211808
algeria	2	14.61880801	0.004504505	0.011507093
bulgaria	2	0.615944778	0.004385965	0.011218762
belarus	2	0	0.004048583	0.005388268
latvia	2	1.129549077	0.004048583	0.005630628
ukraine	2	0.320266314	0.004098361	0.005971287
estonia	2	1.656229935	0.004524887	0.011296919
tanzania	2	40.25432538	0.004524887	0.01150253
luxembourg	2	0.753174215	0.004347826	0.005379732
cyprus	2	0.614103717	0.004444444	0.009508335
jordan	2	4.724074683	0.004237288	0.004073996
namibia	2	3.428027057	0.004484305	0.012541856
zimbabwe	2	0.93482081	0.003717472	0.002365322
brazil	3	233.2845814	0.00591716	0.013924462
japan	3	89.76699793	0.005555556	0.011398148
mexico	3	124.5724364	0.005263158	0.012680571
argentina	3	33.22486757	0.005076142	0.007823186
portugal	3	129.6674323	0.005555556	0.009329678
ireland	3	80.88246759	0.005291005	0.009494981
colombia	3	20.41928805	0.004926108	0.008956424
slovenia	3	3.451929475	0.004608295	0.006652732
serbia	3	2.532112481	0.004273504	0.004720714
uruguay	3	25.28573187	0.004761905	0.007323855
ecuador	3	3.751256425	0.004739336	0.00844905
venezuela	3	0.162799996	0.004081633	0.006566581
paraguay	3	0.670069323	0.004347826	0.004895901
panama	3	0.296590328	0.004273504	0.003678163
cuba	3	0.454435569	0.004310345	0.004887282
oman	3	1.071987111	0.004366812	0.003201285
mozambique	3	0	0.004166667	0.003303245
uzbekistan	3	0.294865068	0.004065041	0.002297688
australia	4	535.2403906	0.006060606	0.013408843
israel	4	52.12090802	0.005181347	0.007577669
egypt	4	37.91583906	0.005128205	0.01119519
malaysia	4	20.93946778	0.004926108	0.008577282
saudi arabia	4	73.08265636	0.005405405	0.012338093
singapore	4	9.684198401	0.004854369	0.006405688
greece	4	32.58605868	0.004854369	0.008536004

romania	4	1.492168025	0.004405286	0.005040863
indonesia	4	0.887149066	0.004291845	0.005571117
lebanon	4	0.91881165	0.004464286	0.004508406
u arab emirates	4	4.223838738	0.004484305	0.008735166
peru	4	4.128752344	0.00462963	0.007002783
qatar	4	0.509585167	0.004424779	0.003058989
sudan	4	3.132404769	0.004484305	0.004011521
nepal	4	2.164760015	0.004504505	0.005836826
kuwait	4	0.145738969	0.004310345	0.004120516
macedonia	4	0.171102652	0.004166667	0.002272918

Table S2. The basic information of the collected high-quality genomes of *Candida albicans*

species	strain_name	assembly_level	scaffolds_number	Genome_size	GC_percentage	N50	N90
Candida albicans	NCYC 4144	Chromosome	8	14699975	0.336045307	2276033	1060689
Candida albicans	NCYC 4145	Chromosome	8	15454853	0.336750345	2451893	1111516
Candida albicans	NCYC 4146	Chromosome	8	15484668	0.3358533	2336946	1236692
Candida albicans	SC5314	Chromosome	8	14282666	0.334555585	2231883	1033292
Candida albicans	TIMM 1768	Chromosome	8	14425299	0.336279928	2264089	1199025
Candida albicans	H892	Contig	13	14578683	0.33548634	1725750	753764
Candida albicans	SC5314-GTH12	Chromosome	17	28593212	0.334839729	2231038	1032182
Candida albicans	SC5314-P0	Chromosome	17	28599939	0.334467016	2231689	1032578
Candida albicans	W0-1	Chromosome	17	14472953	0.334692895	1768732	958737
Candida albicans	ATCC 18804	Chromosome	23	14787852	0.336198547	1642426	1047916
Candida albicans	P76055	Scaffold	38	14453368	0.333685777	988382	365523
Candida albicans	P87	Scaffold	44	14461023	0.334069825	858745	315674
Candida albicans	P37005	Scaffold	45	14473555	0.334353355	769828	251431
Candida albicans	P57072	Scaffold	45	14509911	0.333640224	793722	350221
Candida albicans	P76067	Scaffold	45	14622326	0.333746873	784277	331145
Candida albicans	P60002	Scaffold	47	14791426	0.334308371	1172852	334638
Candida albicans	P37037	Scaffold	48	14475823	0.333675565	473104	288094
Candida albicans	12C	Scaffold	50	14897039	0.334095407	536647	234118
Candida albicans	A203	Scaffold	50	14785429	0.333570689	785987	236111
Candida albicans	P37039	Scaffold	50	14523568	0.333985527	791860	329206
Candida albicans	P75016	Scaffold	50	14680338	0.334454268	1306612	301345
Candida albicans	P75063	Scaffold	51	14453728	0.333797594	741340	304347
Candida albicans	P34048	Scaffold	52	14542907	0.333889347	554132	241195
Candida albicans	A92	Scaffold	54	14607426	0.333555141	763011	273804
Candida albicans	P57055	Scaffold	56	14594318	0.333785314	545440	232682
Candida albicans	P75010	Scaffold	56	14862837	0.334524891	910766	353438
Candida albicans	L26	Scaffold	57	14516741	0.333968908	577645	225852
Candida albicans	P94015	Scaffold	58	14742595	0.33380205	1053458	330295
Candida albicans	P78048	Scaffold	59	14503432	0.333780911	710135	250295
Candida albicans	CA77	Contig	60	14133179	0.336679809	395935	144803
Candida albicans	19F	Scaffold	60	14573765	0.333916823	551228	271191
Candida albicans	A67	Scaffold	64	14688178	0.333699949	581324	244905
Candida albicans	UCSD	Scaffold	66	14641704	0.338402268	322592	153873
Candida albicans	FDAARGOS_656	Contig	69	17026608	0.336127254	833775	58861
Candida albicans	P78042	Scaffold	70	14676842	0.333832939	479789	211261
Candida albicans	A20	Scaffold	71	14553017	0.333876491	810716	237725
Candida albicans	GC75	Scaffold	73	14696915	0.333493296	559034	178724
Candida albicans	A123	Scaffold	75	14638035	0.333402476	601048	168349

Candida albicans	Ca6	Scaffold	82	14716334	0.333517113	400469	184103
Candida albicans	A155	Scaffold	84	14465145	0.333591309	483784	179012
Candida albicans	A84	Scaffold	86	14691637	0.333355546	535335	222519
Candida albicans	A48	Scaffold	87	14699976	0.333415148	422886	139669
Candida albicans	Ca529L	Scaffold	87	14671346	0.334783459	1231433	304407
Candida albicans	ATCC 14053	Scaffold	100	14889980	0.33334147	385709	132990
Candida albicans	ATCC 10231	Scaffold	1404	16602237	0.336747555	30438	5479

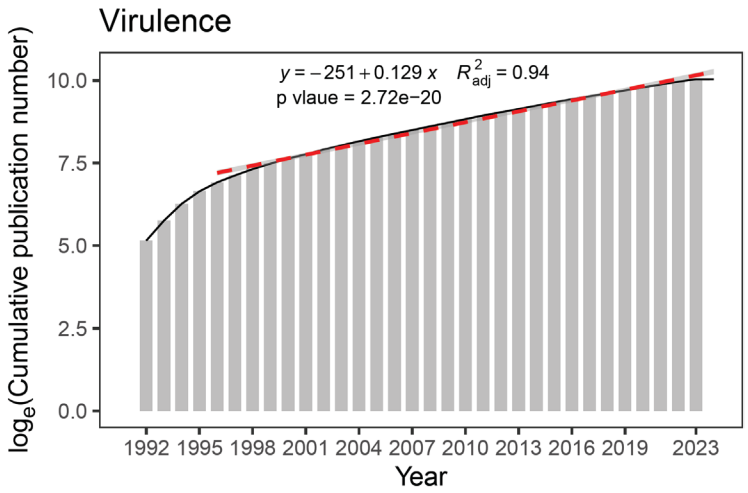


Figure S1. Exponential growth model of publication growth trend.

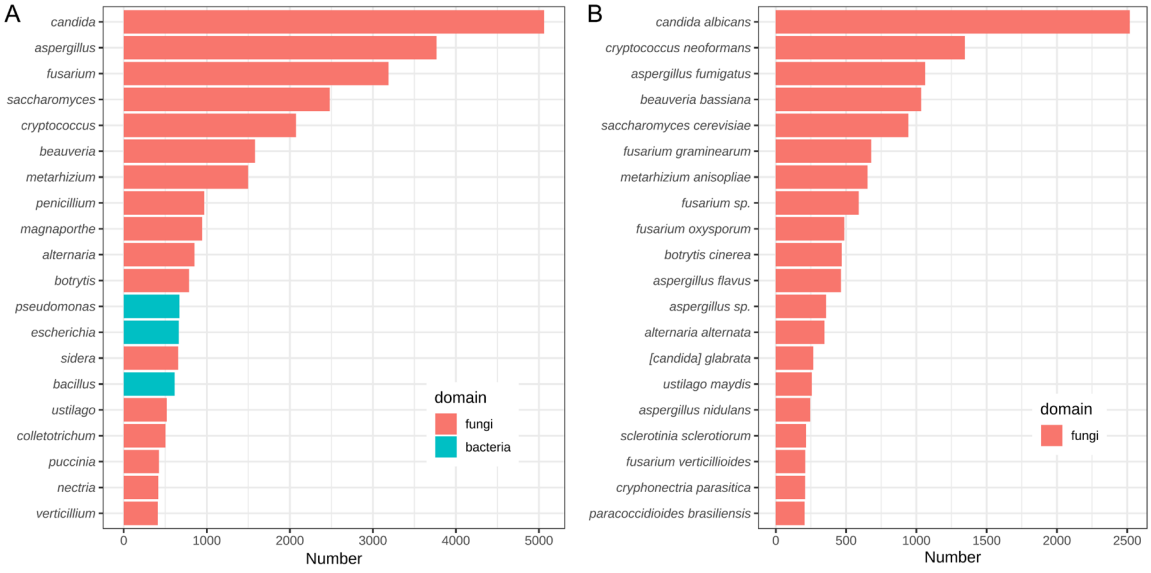


Figure S2. The top 20 genera and species mentioned most frequently in the publications.

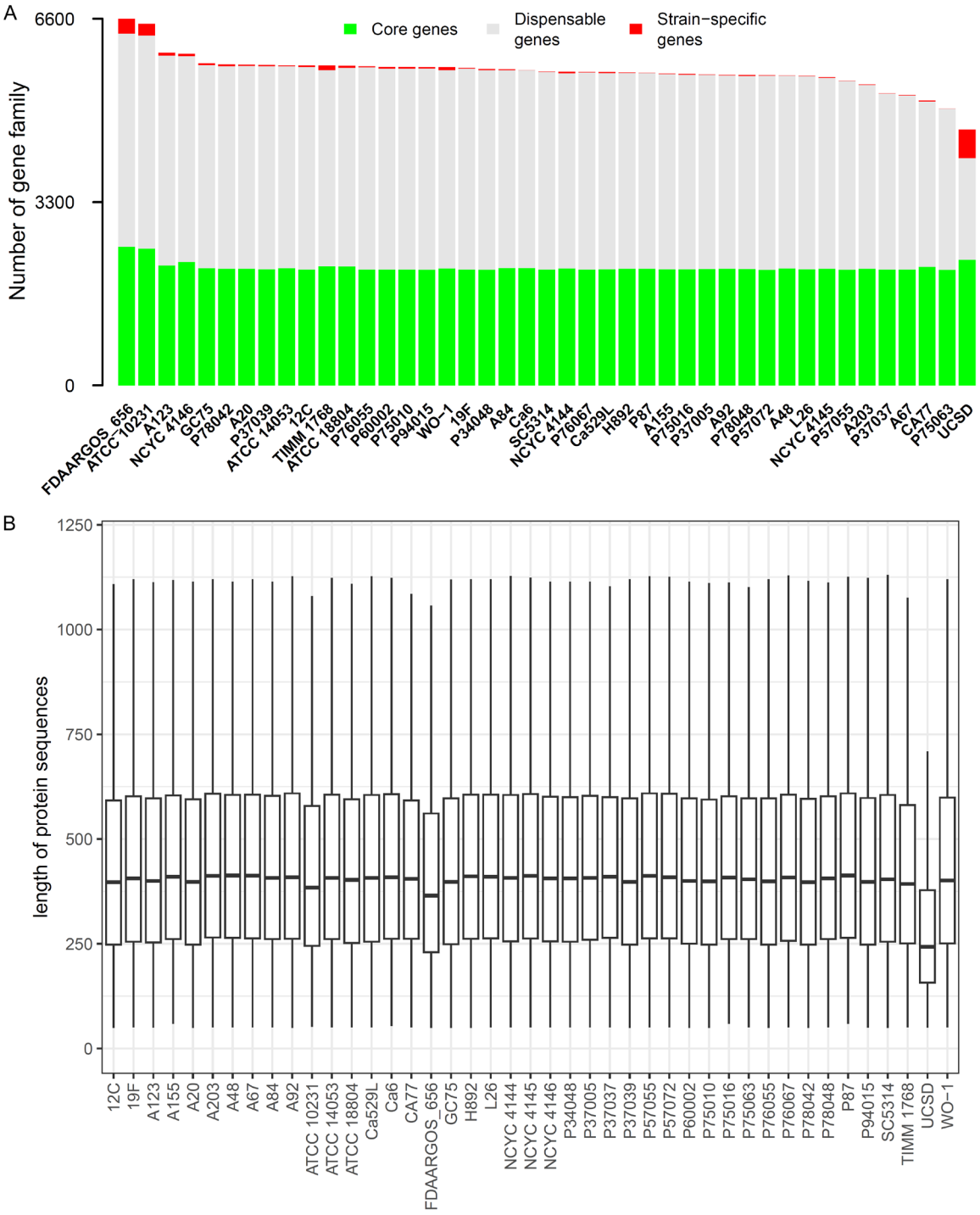


Figure S3. The number and protein length of each strain's gene family.

Table S3. dN/dS value of each sites in ECE1

Partition	Site	ES	EN	S	N	P[S]	dS	dN	dN-dS	P [dN/dS > 1]	P [dN/dS < 1]	Total branch length
1	1	0	3	0	0	0	null	0	null	1	1	0.33133995
1	2	0.584515448	2.213059855	0	0	0.208936448	0	0	0	1	1	0.33133995
1	3	0.652100684	2.347899316	0	0	0.217366895	0	0	0	1	1	0.33133995
1	4	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	5	0.584515448	2.213059855	0	0	0.208936448	0	0	0	1	1	0.33133995
1	6	0.905198178	2.094801822	0	0	0.301732726	0	0	0	1	1	0.33133995
1	7	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	8	0.699746916	2.094801822	0	0	0.250397106	0	0	0	1	1	0.33133995
1	9	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	10	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	11	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	12	0.695205406	2.304794594	2	0	0.231735135	2.876847593	0	-8.682465224	1	0.053701173	0.33133995
1	13	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	14	1.284262364	1.415484552	0	0	0.475697317	0	0	0	1	1	0.33133995
1	15	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	16	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	17	0.584515448	1.763383867	0	0	0.248952519	0	0	0	1	1	0.33133995
1	18	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	19	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	20	0.853621823	2.146378177	0	0	0.284540608	0	0	0	1	1	0.33133995
1	21	0.858839706	2.141160294	4	0	0.286279902	4.657446522	0	-14.05639894	1	0.006716816	0.33133995
1	22	0.652100684	2.347899316	0	0	0.217366895	0	0	0	1	1	0.33133995
1	23	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	24	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	25	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	26	0.584515448	2.303310317	0	0	0.202406757	0	0	0	1	1	0.33133995
1	27	0.652100684	2.347899316	0	0	0.217366895	0	0	0	1	1	0.33133995
1	28	0.652100684	2.347899316	0	0	0.217366895	0	0	0	1	1	0.33133995
1	29	0	3	0	0	0	null	0	null	1	1	0.33133995
1	30	0.692356487	2.105218816	1	0	0.247484486	1.444342644	0	-4.359095981	1	0.247484486	0.33133995
1	31	0.797575304	2	0	0	0.285095205	0	0	0	1	1	0.33133995
1	32	0.665295626	2.301167119	0	3	0.224272369	0	1.303686279	3.934588266	0.466796707	1	0.33133995
1	33	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	34	1	2	1	1	0.333333333	1	0.5	-1.509024191	0.888888889	0.555555556	0.33133995
1	35	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	36	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995

1	37	1	2	1	1	0.333333333	1	0.5	-1.509024191	0.888888889	0.555555556	0.33133995
1	38	0.757778271	1.862056078	0	1	0.289246636	0	0.537040754	1.620814979	0.710753364	1	0.33133995
1	39	1	2	0	1	0.333333333	0	0.5	1.509024191	0.666666667	1	0.33133995
1	40	0.765972892	2.234027108	0	1	0.255324297	0	0.447622142	1.350945282	0.744675703	1	0.33133995
1	41	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	42	1	2	0	0	0.333333333	0	0	0	1	1	0.23897271
1	43	1	2	0	1	0.333333333	0	0.5	2.092289118	0.666666667	1	0.23897271
1	44	0.859002282	2.131817543	1	2	0.287212983	1.164141261	0.938166593	-0.682002481	0.799911251	0.637857627	0.33133995
1	45	0.826737177	1.90132779	0	1	0.303048933	0	0.525948238	1.587337228	0.696951067	1	0.33133995
1	46	0.757778271	2.176825706	0	1	0.258221647	0	0.459384505	1.386444663	0.741778353	1	0.33133995
1	47	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	48	0.342723624	1.033938985	0	0	0.248952519	0	0	0	1	1	0.318476403
1	49	0.999386588	2.000613412	0	2	0.333128863	0	0.999693388	3.017123012	0.444717114	1	0.33133995
1	50	1	2	1	1	0.333333333	1	0.5	-1.509024191	0.888888889	0.555555556	0.33133995
1	51	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	52	1	2	0	1	0.333333333	0	0.5	1.509024191	0.666666667	1	0.33133995
1	53	0.979059048	2.020940952	0	2	0.326353016	0	0.989638019	2.986775422	0.453800259	1	0.33133995
1	54	0.584515448	1.763383867	0	0	0.248952519	0	0	0	1	1	0.33133995
1	55	0.736374592	2.214848203	0.5	4.5	0.24951508	0.679002244	2.031741947	4.082633874	0.435954108	0.761927161	0.33133995
1	56	0.891900798	2.108099202	0	2	0.297300266	0	0.948721957	2.863288766	0.493786916	1	0.33133995
1	57	0.662609879	2.337390121	1	0	0.22086996	1.509183656	0	-4.554789292	1	0.22086996	0.33133995
1	58	1	2	1	0	0.333333333	1	0	-3.018048382	1	0.333333333	0.33133995
1	59	0.935665635	2.064334365	0	1	0.311888545	0	0.484417649	1.461995902	0.688111455	1	0.33133995
1	60	0.925511091	2.074488909	0	1	0.308503697	0	0.482046443	1.454839488	0.691496303	1	0.33133995
1	61	1	2	0	1	0.333333333	0	0.5	1.509024191	0.666666667	1	0.33133995
1	62	0.584515448	2.213059855	0	0	0.208936448	0	0	0	1	1	0.33133995
1	63	0.797575304	2	0	0	0.285095205	0	0	0	1	1	0.33133995
1	64	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	65	0.861176726	2.138823274	0	1	0.287058909	0	0.467546811	1.411078895	0.712941091	1	0.33133995
1	66	0.925511091	2.074488909	0	1	0.308503697	0	0.482046443	1.454839488	0.691496303	1	0.33133995
1	67	0.835712371	2.11032983	2	6	0.283672912	2.393167877	2.84315746	1.358090335	0.59336523	0.71104694	0.33133995
1	68	0.905198178	2.094801822	0	0	0.301732726	0	0	0	1	1	0.33133995
1	69	0.875988232	2.124011768	4	0	0.291996077	4.56627139	0	-13.78122798	1	0.007269559	0.33133995
1	70	0.232259378	2.767740622	0.5	1.5	0.077419793	2.152765605	0.541958299	-4.861494385	0.922580207	0.148845761	0.33133995
1	71	1	2	0	1	0.333333333	0	0.5	1.509024191	0.666666667	1	0.33133995
1	72	0.939881424	2.060118576	0	3	0.313293808	0	1.456226857	4.394963111	0.323826876	1	0.33133995
1	73	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	74	0.943568793	2.032395956	1	6	0.317063162	1.05980614	2.952180643	5.711277807	0.294467491	0.930711082	0.33133995

1	75	0.652100684	2.347899316	0	0	0.217366895	0	0	0	1	1	0.33133995
1	76	0.891724856	2.108275144	2	0	0.297241619	2.242844288	0	-6.769012573	1	0.08835258	0.33133995
1	77	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	78	0.584515448	1.763383867	0	0	0.248952519	0	0	0	1	1	0.33133995
1	79	0.874757556	2.125242444	2	1	0.291585852	2.286347785	0.470534551	-5.480212191	0.975208698	0.205484322	0.33133995
1	80	0.853621823	2.146378177	0	0	0.284540608	0	0	0	1	1	0.33133995
1	81	0.584515448	1.763383867	0	0	0.248952519	0	0	0	1	1	0.33133995
1	82	0.853621823	2.146378177	0	0	0.284540608	0	0	0	1	1	0.33133995
1	83	0.867095144	2.132904856	2	0	0.289031715	2.306551956	0	-6.961285398	1	0.083539332	0.33133995
1	84	0	3	0	0	0	null	0	null	1	1	0.33133995
1	85	0.743983102	2.256016898	3	1	0.247994367	4.032349651	0.443259091	-10.83204896	0.996217602	0.049660616	0.33133995
1	86	0.905198178	2.094801822	0	0	0.301732726	0	0	0	1	1	0.33133995
1	87	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	88	0.640173199	2.157402105	0	2	0.228831445	0	0.927040905	2.797854303	0.594700941	1	0.33133995
1	89	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	90	0.652100684	2.347899316	0	0	0.217366895	0	0	0	1	1	0.33133995
1	91	0.660849549	2.136725755	2	1	0.236222256	3.026407451	0.468005778	-7.72138003	0.986818573	0.141040008	0.33133995
1	92	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	93	0.652100684	2.347899316	0	0	0.217366895	0	0	0	1	1	0.33133995
1	94	0.701968688	2.095606615	0	0	0.250920391	0	0	0	1	1	0.33133995
1	95	0.797575304	2	0	0	0.285095205	0	0	0	1	1	0.33133995
1	96	0.584515448	2.303310317	0	0	0.202406757	0	0	0	1	1	0.33133995
1	97	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	98	0.905198178	2.094801822	0	0	0.301732726	0	0	0	1	1	0.33133995
1	99	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	100	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	101	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	102	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	103	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	104	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	105	0.853621823	2.146378177	0	0	0.284540608	0	0	0	1	1	0.33133995
1	106	0.905198178	2.094801822	0	0	0.301732726	0	0	0	1	1	0.33133995
1	107	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	108	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	109	0	3	0	0	0	null	0	null	1	1	0.33133995
1	110	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	111	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	112	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995

1	113	1	2	4	4	0.333333333	4	2	-6.036096764	0.912056089	0.258649596	0.33133995
1	114	0.797575304	2	0	0	0.285095205	0	0	0	1	1	0.33133995
1	115	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	116	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	117	0.655999982	2.344000018	1	0	0.218666661	1.524390286	0	-4.600683636	1	0.218666661	0.33133995
1	118	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	119	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	120	0.106022665	2.893977335	0	3	0.035340888	0	1.036635624	3.128616467	0.89768013	1	0.33133995
1	121	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	122	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	123	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	124	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	125	1	2	5	0	0.333333333	5	0	-15.09024191	1	0.004115226	0.33133995
1	126	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	127	0.701968688	2.095606615	0	0	0.250920391	0	0	0	1	1	0.33133995
1	128	0.797575304	2	0	0	0.285095205	0	0	0	1	1	0.33133995
1	129	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	130	1	1.887825765	0	0	0.346281279	0	0	0	1	1	0.33133995
1	131	1	2	0	5	0.333333333	0	2.5	7.545120955	0.131687243	1	0.33133995
1	132	0.659382903	2.340617097	5	0	0.219794301	7.582847507	0	-22.88540065	1	0.000512958	0.33133995
1	133	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	134	0.655999982	2.344000018	1	0	0.218666661	1.524390286	0	-4.600683636	1	0.218666661	0.33133995
1	135	0.991654816	2.008345184	0	2	0.330551605	0	0.995844746	3.005507625	0.448161153	1	0.33133995
1	136	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	137	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	138	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	139	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	140	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	141	0.797575304	2	0	0	0.285095205	0	0	0	1	1	0.33133995
1	142	1.401715604	1.392833134	0	0	0.501589249	0	0	0	1	1	0.33133995
1	143	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	144	0.584515448	2.303310317	0	0	0.202406757	0	0	0	1	1	0.33133995
1	145	0.905198178	2.094801822	0	0	0.301732726	0	0	0	1	1	0.33133995
1	146	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	147	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	148	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	149	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	150	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995

1	151	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	152	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	153	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	154	0.584515448	1.763383867	0	0	0.248952519	0	0	0	1	1	0.33133995
1	155	0.584515448	1.763383867	0	0	0.248952519	0	0	0	1	1	0.33133995
1	156	1	2	2	0	0.333333333	2	0	-6.036096764	1	0.111111111	0.33133995
1	157	0.940496341	2.059503659	1	3	0.31349878	1.063268357	1.456661651	1.187279994	0.627822343	0.777891463	0.33133995
1	158	0.584515448	2.303310317	0	0	0.202406757	0	0	0	1	1	0.33133995
1	159	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	160	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	161	0.701968688	2.095606615	0	0	0.250920391	0	0	0	1	1	0.33133995
1	162	0.797575304	2	0	0	0.285095205	0	0	0	1	1	0.33133995
1	163	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	164	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	165	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	166	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	167	0.65900098	2.34099902	2	0	0.219666993	3.034896853	0	-9.159465537	1	0.048253588	0.33133995
1	168	0.945558179	2.054441821	0	3	0.31518606	0	1.460250647	4.407107104	0.321157285	1	0.33133995
1	169	1	2	3	0	0.333333333	3	0	-9.054145146	1	0.037037037	0.33133995
1	170	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	171	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	172	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	173	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	174	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	175	0.695847618	2.304152382	1	0	0.231949206	1.437096247	0	-4.337226003	1	0.231949206	0.33133995
1	176	1	2	1	0	0.333333333	1	0	-3.018048382	1	0.333333333	0.33133995
1	177	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	178	0.797575304	2	0	0	0.285095205	0	0	0	1	1	0.33133995
1	179	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	180	0.706725606	2.293274394	0	5	0.235575202	0	2.180288592	6.580216457	0.261020031	1	0.33133995
1	181	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	182	0.659469995	2.340530005	4	4	0.219823332	6.065476866	1.709014621	-13.14801383	0.984266605	0.076290488	0.33133995
1	183	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	184	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	185	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	186	1	2	1	0	0.333333333	1	0	-3.018048382	1	0.333333333	0.33133995
1	187	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	188	1.652100684	1.347899316	0	0	0.550700228	0	0	0	1	1	0.33133995

1	189	0.584515448	2.303310317	0	0	0.202406757	0	0	0	1	1	0.33133995
1	190	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	191	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	192	0.594127649	1.79057249	1	1	0.24914145	1.68313998	0.558480601	-3.394276421	0.937928538	0.436211437	0.33133995
1	193	0.584515448	1.763383867	0	0	0.248952519	0	0	0	1	1	0.33133995
1	194	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	195	0.692356487	2.105218816	1	0	0.247484486	1.444342644	0	-4.359095981	1	0.247484486	0.33133995
1	196	0.797575304	2	0	0	0.285095205	0	0	0	1	1	0.33133995
1	197	0.690316541	2.300503284	0	1	0.230811811	0	0.434687491	1.311907878	0.769188189	1	0.33133995
1	198	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	199	0.974626412	2.025373588	0	2	0.324875471	0	0.987472145	2.980238708	0.45579313	1	0.33133995
1	200	0.584958304	2.302867461	1	0	0.20256011	1.709523555	0	-5.159424799	1	0.20256011	0.33133995
1	201	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	202	0.952368227	2.047631773	0	6	0.317456076	0	2.930214347	8.843528671	0.10110756	1	0.33133995
1	203	0.9921103	2.0078897	0	2	0.330703433	0	0.996070651	3.006189416	0.447957894	1	0.33133995
1	204	0.626215671	2.08128105	1	10	0.23128954	1.596893924	4.804733124	9.681413906	0.238690773	0.944615186	0.33133995
1	205	0.584515448	2.093570715	0	4	0.218258642	0	1.910611364	5.766317537	0.373467106	1	0.33133995
1	206	1	2	4	0	0.333333333	4	0	-12.07219353	1	0.012345679	0.33133995
1	207	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	208	0.64535402	2.162605542	2	8	0.22983024	3.099074213	3.699241421	1.811333674	0.586801578	0.707449833	0.33133995
1	209	0.646792339	1.842301687	1.5	5.5	0.259850505	2.31913693	2.985395953	2.010801968	0.578213938	0.728792157	0.33133995
1	210	1	2	1	0	0.333333333	1	0	-3.018048382	1	0.333333333	0.33133995
1	211	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	212	0.589229136	1.911869108	0	5	0.235588161	0	2.615241796	7.892926271	0.260997906	1	0.33133995
1	213	1	2	2	0	0.333333333	2	0	-6.036096764	1	0.111111111	0.33133995
1	214	0.901601019	2.098398981	1	0	0.300533673	1.109138054	0	-3.347432311	1	0.300533673	0.33133995
1	215	0.979059048	2.020940952	0	5	0.326353016	0	2.474095047	7.466938554	0.138727273	1	0.33133995
1	216	0.81169328	1.992176482	0	1	0.289490365	0	0.50196356	1.514950311	0.710509635	1	0.33133995
1	217	1	1.68383175	3	1	0.372601598	3	0.593883564	-7.261775818	0.980725689	0.149093096	0.33133995
1	218	0.985662047	2.014337953	1	2	0.328554016	1.01454652	0.992882052	-0.065384413	0.747090104	0.697285487	0.33133995
1	219	0.62563945	2.270009833	1.5	1.5	0.216061888	2.397547022	0.660790089	-5.241616449	0.935019071	0.319049651	0.33133995
1	220	0.979059048	2.014337953	0	2	0.327072903	0	0.992882052	2.996566071	0.452830877	1	0.33133995
1	221	0.975427728	2.024572272	0	3	0.325142576	0	1.481794471	4.472127407	0.307352033	1	0.33133995
1	222	1.401715604	1.392833134	0	0	0.501589249	0	0	0	1	1	0.33133995
1	223	0.599375167	2.198200137	0	1	0.21424809	0	0.454917631	1.372963422	0.78575191	1	0.33133995
1	224	0.652100684	2.347899316	0	1	0.217366895	0	0.425912642	1.285424959	0.782633105	1	0.33133995
1	225	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	226	0.584515448	1.763383867	0	0	0.248952519	0	0	0	1	1	0.33133995

1	227	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	228	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	229	0.584515448	2.213059855	0	0	0.208936448	0	0	0	1	1	0.33133995
1	230	0.797575304	2	0	0	0.285095205	0	0	0	1	1	0.33133995
1	231	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	232	1	2	4	0	0.333333333	4	0	-12.07219353	1	0.012345679	0.33133995
1	233	1	2	1	0	0.333333333	1	0	-3.018048382	1	0.333333333	0.33133995
1	234	1	1.652100684	0	0	0.377059591	0	0	0	1	1	0.33133995
1	235	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	236	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	237	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	238	0.652100684	2.347899316	0	0	0.217366895	0	0	0	1	1	0.33133995
1	239	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	240	1.289832984	1.414410231	1	0	0.476966338	0.775294176	0	-2.339875333	1	0.476966338	0.33133995
1	241	0.855557657	2.144442343	2	0	0.285185886	2.337656597	0	-7.055160711	1	0.081330989	0.33133995
1	242	0.586958941	2.210616363	1	0	0.209809881	1.703696682	0	-5.141839015	1	0.209809881	0.33133995
1	243	0.699525196	2.098050108	1	0	0.250046959	1.429541074	0	-4.314424125	1	0.250046959	0.33133995
1	244	0.691151986	2.308848014	3	0	0.230383995	4.340579293	0	-13.10007831	1	0.012228042	0.33133995
1	245	0.905198178	2.094801822	0	0	0.301732726	0	0	0	1	1	0.33133995
1	246	0.584515448	2.303310317	0	0	0.202406757	0	0	0	1	1	0.33133995
1	247	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	248	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	249	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	250	0.906784155	2.093215845	0	1	0.302261385	0	0.477733819	1.44182378	0.697738615	1	0.33133995
1	251	0.584515448	2.303310317	0	0	0.202406757	0	0	0	1	1	0.33133995
1	252	0.699746916	2.300253084	0	1	0.233248972	0	0.434734772	1.312050575	0.766751028	1	0.33133995
1	253	1	2	0	1	0.333333333	0	0.5	1.509024191	0.666666667	1	0.33133995
1	254	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	255	0.584515448	1.763383867	0	0	0.248952519	0	0	0	1	1	0.33133995
1	256	0.762380878	2.237619122	2	0	0.254126959	2.623360654	0	-7.917429379	1	0.064580512	0.33133995
1	257	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995
1	258	1	1.652100684	0	0	0.377059591	0	0	0	1	1	0.33133995
1	259	0.584515448	2.296951661	0	2	0.202853417	0	0.870719238	2.627872788	0.635442675	1	0.33133995
1	260	0.797575304	2	0	0	0.285095205	0	0	0	1	1	0.33133995
1	261	0.699746916	2.300253084	0	0	0.233248972	0	0	0	1	1	0.33133995
1	262	0.905198178	2.094801822	0	0	0.301732726	0	0	0	1	1	0.33133995
1	263	1	2	0	3	0.333333333	0	1.5	4.527072573	0.296296296	1	0.33133995
1	264	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995

1	265	1.371616348	1.398637919	3	0	0.495122908	2.187200527	0	-6.601077011	1	0.121377744	0.33133995
1	266	0.905198178	2.094801822	0	0	0.301732726	0	0	0	1	1	0.33133995
1	267	0.584515448	2.303310317	0	0	0.202406757	0	0	0	1	1	0.33133995
1	268	0.586068624	2.303269109	0	1	0.202838394	0	0.434165507	1.310332505	0.797161606	1	0.33133995
1	269	0.699746916	2	0	0	0.259189819	0	0	0	1	1	0.33133995
1	270	0.652100684	2.347899316	0	0	0.217366895	0	0	0	1	1	0.33133995
1	271	1	1.887825765	0	0	0.346281279	0	0	0	1	1	0.33133995
1	272	0.699527895	2.300087356	0	2	0.233205874	0	0.869532192	2.624290224	0.587973232	1	0.33133995
1	273	1	2	0	0	0.333333333	0	0	0	1	1	0.33133995

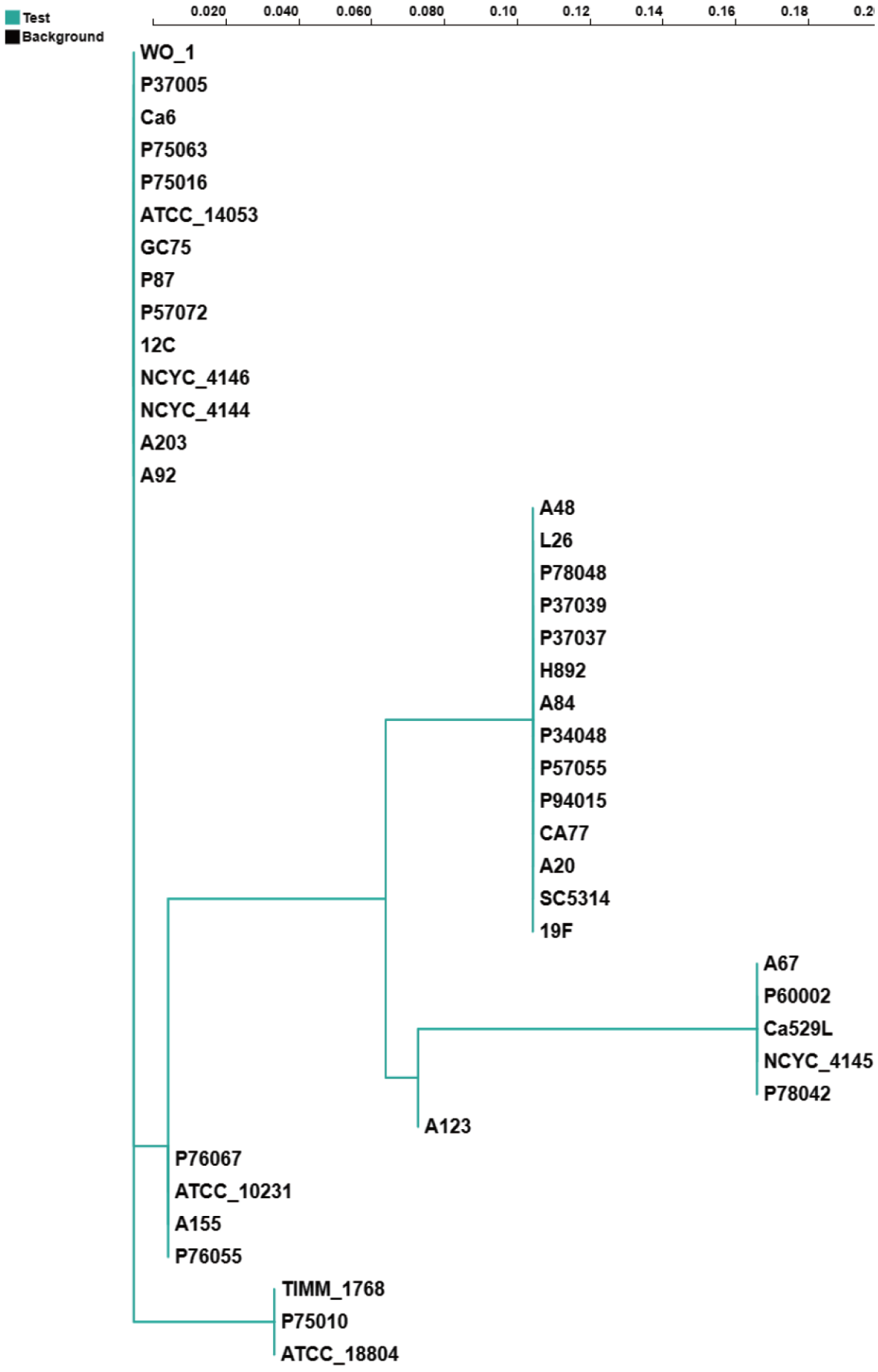

Test
Background
0.020
0.040
0.060
0.080
0.10
0.12
0.14
0.16
0.18
WO_1
P37005
Ca6
P75063
P75016
ATCC_14053
GC75
P87
P57072
12C
NCYC_4146
NCYC_4144
A203
A92
A48
L26
P78048
P37039
P37037
H892
A84
P34048
P57055
P94015
CA77
A20
SC5314
19F
A67
P60002
Ca529L
NCYC_4145
P78042
A123
P76067
ATCC_10231
A155
P76055
TIMM_1768
P75010
ATCC_18804

Figure S4. Evolutionary tree based on dN/dS analysis of ECE1-III.

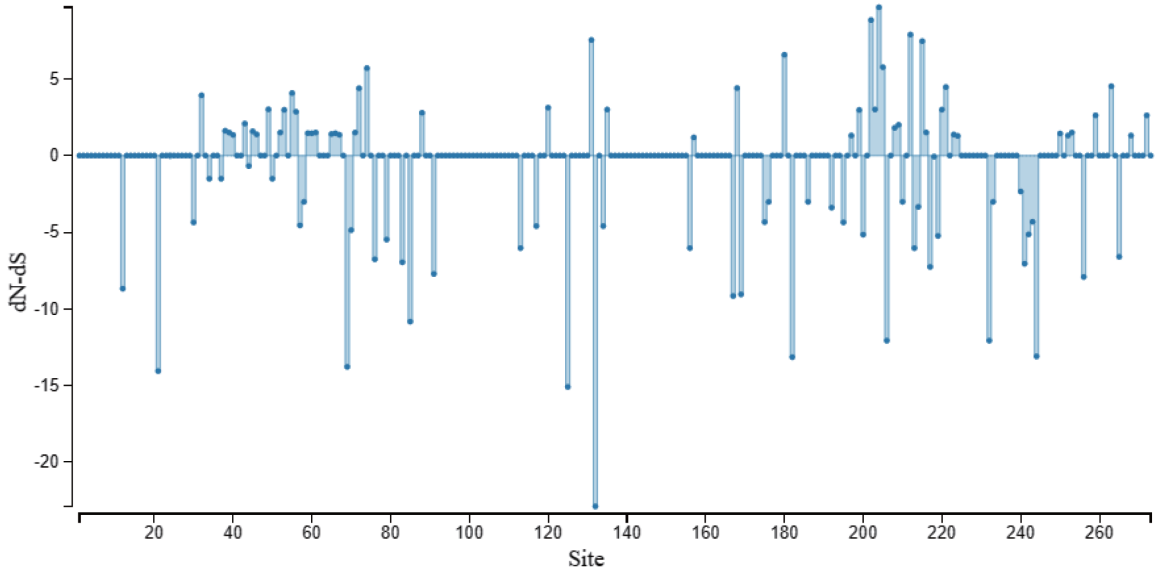


Figure S5. Clustering and binding energy of top 100 candidate drugs.