Hybrid genome Sequence of *Cryptococcus neoformans* of Indian origin and comparative genome analysis

Jananishree Sathiyamoorthy[1] and Jayapradha Ramakrishnan[*]

First Author: **Jananishree Sathiyamoorthy[1]**

Actinomycetes Bioprospecting Lab

Centre for Research in Infectious Diseases (CRID),

School of Chemical and Biotechnology, SASTRA Deemed University,

Tirumalaisamudram, Thanjavur – 613401, Tamilnadu, India.

Email: jananishreegsm96@gmail.com

Corresponding Author: **Jayapradha Ramakrishnan[*]**

Actinomycetes Bioprospecting Lab

Centre for Research in Infectious Diseases (CRID),

School of Chemical and Biotechnology, SASTRA Deemed University,

Tirumalaisamudram, Thanjavur – 613401, Tamilnadu, India.

Email: kavijayashal@gmail.com; antibioticbiology@gmail.com

Ph: +91-9952538211

Abstract

The human opportunistic fungi C. neoformans is a global health concern as it causes severe pneumonia and cryptococcal meningitis. Though the occurrence of the infection has reduced from 2014 to 2022 still the mortality rate remains high and is considered to be a neglected tropical disease. Besides the knowledge about the epidemiological occurrence and sequence type identification of Cryptococcus in different parts of the world, more studies are needed to understand the complete genome, validating the genomic profiles and pan-genome to underpin the disease management. The complete genome profile of 12 strains of Cryptococcus neoformans was analyzed in the study for the virulome, resistome and pangenome. The subspecies of the lab strain were identified as grubii. From the study conducted it was identified that the strains Cn, H99 and JEC21 are more virulent because of the presence of all the necessary virulence genes. The virulence pattern differed in the case of melanization, evasion of host immune response, transcriptional regulation and oxidative stress response. The evolutionary relatedness study shows no unique gene expansion and contraction between the study strains indicating the presence of stable gene families or equal evolutionary changes. There is about 74% of orthologous gene clusters are commonly shared indicating the high degree of conserved functional elements across the species. To the best of our knowledge, this was the first genomic study of Cryptococcus sp. to study the virulome, resistome and pan-genome profiles. Our study highlights the necessity of bioinformatic tools to efficiently study the complete genome and to carry out comparative genome analysis on a large scale.

Keywords: *Cryptococcus neoformans* sp., Comparative genomics, Phylogeny, Virulome, Resistome, Pan-genome

1. Introduction

Cryptococcus sp. is a dimorphic environmental fungus that majorly infects HIV patients (do Carmo et al., 2022). It is considered intricate because it is classified into two broad species, *C. neoformans* and *C. gatti*. Furthermore, two subspecies, *C. neoformans* var. *grubii* (serotype A) and *C. neoformans var. neoformans* (serotype D) were found. *C. gatti* was classified into serotype B and serotype C. Currently there are nine major molecular types (Kwon-chung et al., 2014). Among these, serotype A (*C. neoformans var. grubii*) is highly pathogenic and the molecular types VNI and VNII are widespread in the Indian population (Xess et al., 2021). *C. neoformans* is considered one of the critical priority pathogens by WHO (Casalini et al., 2024; WHO, 2021). The high pathogenicity of serotype A is due to its escalated virulence factors, environmental adaptability, and immune evasion capabilities (Chayakulkeeree and Perfect, 2006; Sabiiti and May, 2012).

This ubiquitous pathogen frequently inhabits avian environments (pigeon droppings, claws and wings of pigeons), marshy areas, and decaying wood (Coelho et al., 2014; Firacative et al., 2021). The infection begins in the lungs through the inhalation of cryptococcal spores/ infectious propagules. Once *Cryptococcus* colonizes the lungs it breaches the lungs and circulates to the liver, kidney and bone marrow through the bloodstream and disseminates to the brain via paracytosis, transcytosis and trojan horse mechanism (Maziarz and Perfect, 2016; Rathore et al., 2022). The *Cryptococcus* sp. exhibits neurotropism owing to the presence of proteinases and urease that allow them to penetrate the blood-brain barrier (degradation of ECM components of the host) and survive. Tropical, subtropical and temperate locations have been found to have higher incidences of cryptococcal infection (Diniz-lima et al., 2022). The prevalence of cryptococcal infection was high majorly in Sub-Saharan Africa (incidence: 6-10% among individuals with advanced HIV disease; mortality: 24% to 47% at 10 weeks post-diagnosis), followed by Asia and Pacific (incidence: 1 to 3% among the HIV infected individuals) (Rajasingham et al., 2022). These areas are influenced by a combination of environmental, ecological, socioeconomic, and healthcare-related factors. It was estimated that 1, 52, 000 new cases and 1, 12, 000 (19%) deaths were reported globally each year. The advent of HAART reduced the incidence of cryptococcal infection but still the mortality rate remains high (Rajasingham et al., 2022).

In the recent decade it was noted that immune-competent individuals are also acquiring the infection and the reason could be the result of the organism's increased pathogenicity, which

could be caused by virulence, stress tolerance, or AntiFungal Resistance (AFR) (Maziarz and Perfect, 2016). The virulence factors of *Cryptococcus* sp. are the capsular polysaccharides (glucoronoxylomannan, galactoxylomannan), melanin, mannoproteins and certain lytic enzymes (urease, phospholipase and proteases) (Chen et al., 2022; Snelders et al., 2022).

The preliminary role of the stress tolerance genes was to overcome the environmental stressors (temperature, pH, reactive oxygen species, reactive nitrogen species and nutrient starvation). In addition, the autophagy genes (also stress-tolerant genes) in *Cryptococcus* sp., play a role in cell differentiation, cell growth, morphology and pathogenesis. These ATGs play a critical role in bilateral reproduction (Ding et al., 2018; Jiang et al., 2020; Zhao et al., 2019). The recent insights of cryptococcal genomics reveal that stress-tolerant genes play a pivotal role in modulating the major virulence factors synthesis and promoting AFR (Brown et al., 2007). They also regulate the genes responsible for maintaining the cell wall, cell membrane integrity and also regulate the evasion of host immune response.

The development of AFR or tolerance, which is brought about by stable and heritable point mutations that impact the target and drug interaction is significantly influenced by the organism's genetic background (Gow et al., 2022). Complete Genome Sequencing (CGS) has revolutionized the understanding of AFR in *C. neoformans.* Though the inherent resistance (echinocandins resistance) is more common in *Cryptococcus* sp. the fungal genomics highlights the insights of acquired resistant mechanisms (target protein modification, over-expression of target proteins, and upregulation of multi-drug transporters) caused by genetic mutations and structural variations (Cao et al., 2019; Gow et al., 2022; Park et al., 2016). The mutation's non-primary target proteins also produce resistance to antifungal drugs (Alastruey-Izquierdo and Martín-Galiano, 2023)

In the earlier decades, PCR-Restriction Fragment Length Polymorphism and PCR fingerprinting were used to study *Cryptococcus* molecular epidemiology, the genetic structure of the fungus and genome plasticity. The recent advent and advancements in WGS help in understanding the epidemiology and genetic structure of the *Cryptococcus* sp. (Patiño et al., 2023). Globally the studies based on the WGS have demonstrated small genetic variability and a small number of chromosomal rearrangements across the molecular types and also the variations in gene content and structural changes due to translocations, aneuploidy and inversions were noted in *C. neoformans*. The WGS also helps in revealing the hybrids present between the *Cryptococcus* sp., complex (Patiño et al., 2023). In addition, WGS helps in

identifying the microevolution signals of infecting strains revealing the rapid adaptation of the organism through hypermutation to identify genes related to the adaptation of an organism to the stress condition that include virulence and pathogenicity features and monitoring of outbreaks (Arastehfar et al., 2020; Patiño et al., 2023).

Considering it as a critical priority pathogen as described in the guidelines by WHO and the gap in the knowledge about the genomic features of clinical *Cryptococcus* species of Indian origin, this was the first study in India we demonstrated hybrid sequencing of *Cryptococcus neoformans* (clinical strain) and unravel the genomic aspects and comparative pan-genome analyses were made with the clinical isolates and an environmental isolate from NCBI datasets to understand the evolutionary relationships, resistome and virulome. To the best of our knowledge, this is the first kind of study to perform the hybrid sequence of the *Cryptococcus* genome, its resistome, virulome and pan-genome analysis.

2. Materials and Methods

2.1. Culture Isolation and DNA extraction

The comprehensive methodology is depicted in Figure 1. The *Cryptococcus* sp. (Cn) was collected from the microbiology lab of Trichy Medical College, Tamil Nadu, India sourced from the cryptococcal meningitis patient. The presumptive identification was performed by Indian ink staining after growing the yeast colonies in potato dextrose agar containing streptomycin. The purified yeast colonies were stored in 30% glycerol at -80°C.

Initially, the clinical isolate Cn was grown in Potato Dextrose Broth (PDB) and incubated for 48 hours at 37°C. Approximately 10^7 log phase cells were obtained by centrifugation at 7500 rpm for 10 minutes following the scheduled incubation period. The isopropanol precipitation technique, phenol-chloroform, and urea lysis buffer were used to extract the DNA from the pellet. After two cycles of mechanical lysis, one cycle was finished with one milliliter of urea lysis buffer and four hours of incubation at 30°C. Proteinase K (0.02 ml) and Sarkosyl lysis buffer (0.4 ml) were added to the lysate, and the mixture was incubated for 1 hour at 55°C. After a three-minute break on ice and another three minutes of mechanical lysis at 2500 rpm, the cycle was repeated three times. RNase A was added to the lysate, and it was then incubated for 20 minutes at 65°C. The lysate was blended with an equal volume of phenol-chloroform, inverted, and centrifuged for 10 minutes at 4°C and 13,200 rpm. After transferring the aqueous phase to a fresh microcentrifuge tube and adding an equivalent volume

of chloroform, isoamyl alcohol was added and the tube was centrifuged. After transferring the aqueous phase to a fresh microcentrifuge tube, DNA was precipitated in an equivalent volume of isopropanol. The pellet was washed with 70% ethanol, air dried and dissolved in 10mM Tris-Cl (pH 8.0).

The concentration and purity of genomic DNA were quantified using a Nanodrop Spectrophotometer (Thermo Scientific; 2000). DNA integrity was observed on agarose gel electrophoresis (supplementary figure 1). DNA concentration quantified using Qubit dsDNA HS assay kit.

2.2. Complete Genome Sequencing

2.2.1 Illumina library preparation, sequencing and preprocessing

The library preparation was performed with 50ng of quantified DNA. The QIASeq FX DNA Library Kit was used to prepare the library according to the protocol provided in the kit manual ("QIAseq ® FX DNA Library Kit Handbook," 2023). Indexing PCR was performed on the library to enrich the adapter-tagged fragments and purification of the amplified library was performed with High Prep TM beads. The library was quantified using a Qubit fluorometer (Thermo Fisher Scientific, MA, USA), and the Agilent 2200 Tape Station was used to evaluate the fragment size distribution of the library (supplementary figure 2).

The libraries were sequenced on Illumina NovaSeq 6000 (Illumina, San Diego, USA) using 150 bp paired-end chemistry. The data obtained from the sequencing run was de-multiplexed using Bcl2fastq2 softwarev2.20 and FastQ files were generated based on the distinct dual barcode sequences. The sequencing quality was assessed using FastQC3 v0.11.8 software and Trimgalore- v0.4.0.

2.2.2. Nanopore Library preparation, Sequencing and preprocessing of reads

The NEBnext ultra II end repair kit was used to end polish and A-tail the genomic DNA. A blunt TA ligase master mix was used to barcode the end-prepared samples. With NEB Quick T4 DNA Ligase, sequencing adapters were ligated onto double standard DNA fragments to create the library. The prepared library was ready to use the following purification using AMPure XP beads. Qubit was used to quantify the prepared library.

The estimated DNA was combined and sequenced using a PromethION flow cell (FLO-PRO114M) on a Nanopore PromethION system (PromethION P24 and Data Acquisition Unit, ONT, Oxford, UK) following the necessary data amount and effective library concentration. Porechop-v0.2.4 was used to process the Nanopore raw reads.

2.2.3. Assembly and Annotation

The high-quality processed reads were used for d*e novo* assembly using the MaSuRCA5-v3.4.2 assembler (Zimin et al., 2013). The assembled genome was further improved by closing the gaps using LRGapcloser6-v3.0 and finally, scaffolds were constructed using the LRScaf-v1.1.12 (Xu et al., 2018). The assembled genome was utilized for gene and protein prediction using the eukaryotic genome prediction tool, Braker-v2.1.4 gene prediction program (Hoff et al., 2021). The predicted proteins were searched against the Uniprot fungal database using DIAMOND-v0.8.29 (Buchfink et al., 2015). The KAAS database was used to retrieve information from the BlastP program for functional annotation and pathways (Moriya et al., 2007). The database uses BLAST comparisons against the manually edited KEGG GENES database to offer functional annotation of genes (Moriya et al., 2007). Along with the annotation of the isolated study strain Cn, 11 more assembled sequences were also annotated by using the tool AUGUSTUS v3.4.0 (Stanke and Morgenstern, 2005)

2.3. Comparative genome analysis with reference genome

The assembled genome was also compared with the reference genome using two web-based resources, the Average Nucleotide Identity (ANI) calculator and D-Genies which are tools used to quickly compare the sequences at the whole genome scale (Ha et al., 2017). The assembled genome was validated for completeness and quality of the genome assembly using BUSCO-v5.4.2 (Manni et al., 2022). The Mauve- v2.4.0 tool was used to compare the draft genome against the closest reference genome for the construction of multiple genomes in the presence of large-scale evolutionary events such as rearrangement and inversion (Mau and Perna, 2014).

2.4. Retrieval of global genome sequence data of *Cryptococcus neoformans*

The WGS of 11 isolates of *C. neoformans* were collected from the NCBI datasets (https//www.ncbi.nlm.nih.gov/datasets/) in FASTA format. The sequences were further analyzed for evolutionary relationship, virulence and AFR. Due to the lack of fungal databases

for virulence and AFR analysis, we have limited our comparative study to 12 isolates (n=11 isolates from NCBI datasets and n=1 hybrid sequenced genome).

2.5. Phylogenetic analysis

The phylogenetic analysis was carried out between the study strain Cn and the retrieved *C. neoformans* from the NCBI datasets. The extracted sequences from NCBI datasets include the species isolated from clinical (n=10) and environment (n=1). To investigate the phylogenetic placement based on MLST, the seven housekeeping genes *CAP59*, *GPD1*, *IGS1*, *SOD1*, *LAC1*, *URA5* and *PLB1* were used. Each genome was concatenated and the phylogenetic tree was constructed by the maximum likelihood method using MegaX software (Kumar et al., 2018). Tree editing and annotation were performed using the interactive Tree of Life (iTOL) (Letunic and Bork, 2024).

2.6. Virulence gene identification and comparative genomic study

The virulence-related genes (VRGs) (n=272) of *C. neoformans* were collected from the literature and DFVF (Database of Fungal Virulence Factors) database (Lu, Yao, and Zhang 2012; Almeida and Wolf 2015; Ding et al. 2018; Jiang et al. 2020; Zaragoza 2019; Brown, Campbell, and Lodge 2007; Malachowski et al. 2016; Zhang et al. 2006; Verma, Shakya, and Idnurm 2018; Brandão et al. 2018; Azevedo, Rizzo, and Rodrigues 2016). The listed gene sequences were retrieved from Uniprot and their presence in the annotated study genomes were identified using BLAST. The sequence with 99 to 100% identity was considered for the presence or absence of VRGs. The comprehensive result of the comparative virulome analysis was consolidated using iTOL.

2.7. AFR gene identification and comparative Analysis of Mutation Profiles

Similarly, the AFR genes of *C. neoformans* were collected from the literature. The AFR gene sequences were retrieved from Uniprot and their presence in the annotated genome was identified using BLAST with the query coverage and identity of 99 to 100% is considered as the presence or absence of virulence genes. Followed by, the protein mutations responsible for the antifungal resistance provided in the AFRbase database were compared with the annotated genomes of the study strains using CLUSTAL Omega. The comprehensive result of the comparative AFR analysis and mutation profiles was consolidated using iTOL.

2.8. Pan-genome analysis

The pan-genome analysis was performed using Orthovenn3 with a FASTA annotation file and GFF file produced by AUGUSTUS (genome annotation tool). The data was analyzed and visualized in Orthovenn3. The orthologous gene analysis was carried out by the OrthoFinder algorithm and the Markov clustering algorithm was used to generate the gene clusters. The evolutionary phylogenetic tree construction was made using FastTree2 by the maximum likelihood method. The gene family contraction and expansion analysis were performed using CAFE5 (Sun et al., 2023).

3. Result

3.1. Sequencing and Assembly

The hybrid technology resulted in a total of 7.35 million reads at 116.05X coverage for Illumina sequencing and a total of 6.09 lakhs raw reads at 104.8X coverage for Nanopore sequencing. The genome sequences of the strain Cn was deposited under BioProject accession number PRJNA650119. The assembled genome was approximately 19 MB in size with 20 contigs. The quality check of the assembled genome showed 89.6% completeness.

3.2. Gene prediction and Annotation

A total of 5673 predicted proteins were annotated successfully using the BRAKER tool. The Gene Ontology annotation categorizes the proteins into 3 groups such as biological process (14.87%), cellular components (41.75%) and molecular function (43.37%). The detailed gene ontology annotation is provided in Figure 2. The functional annotation using KEGG pathways annotated 5963 genes which fall majorly into 6 categories, which include BRITE hierarchy (3186 genes), metabolism (1294 genes), genetic information processing (831 genes), cellular processes (419 genes) and environmental information processing (97 genes). The genes (n=136) of unknown functions are categorized separately. The assembled genome was identified with 29 cellular pathways using the KAAS database. The top 10 functional pathways are depicted in Figure 3.

3.3. Genome identity search of clinical Cn

The genome identity search using the ANI calculator indicated a 99.9% identity with the reference genome (H99). The same result was depicted in a whole genome analysis by MAUVE as a dot plot image and D Genies by a linear pairwise genome comparison image

(figure 4a, 4b and 4c). The comparison of ITS region was shown to have 100% coverage and identity with *C. neoformans* var. *grubii* CBS8710 strain, isolated from Hodgkin's lymphoma patient from North Carolina.

3.4. Phylogenetic analysis

The midpoint was divided into two major clades. Clade 1 represents the environmental strain VNII. Other clinical strains were grouped in clade 2. JEC21 (*C. neoformans* var. *neoformans*) was out-grouped in clade 2 from the other clinical strains of *C. neoformans*. From the phylogenetic tree, it was noted that the study strain Cn was closely aligned with the reference strain H99 (figure 5).

3.5. Virulome and comparative virulome analysis

Due to the lack of virulome profiling databases, the Virulence Related Genes (VRGs) collected from the literature and DFVF database were identified in the 12 genomes using BLAST (>99% coverage and >99% identity). From the literature and DFVF database, 272 VRGs of *C. neoformans* were retrieved. Based on the functionality we categorized the VRGs into stress tolerance (n=164), capsular-associated (n=78), cell wall-associated (n=85), cell membrane-associated (n=24), nutrient acquisition, iron acquisition and homeostasis (n=39), host cell adhesion (n=32), host cell colonization (n=16), evasion of host immune response (n=49), biofilm-associated (n=9), protease (n=7), urease (n=1), toxin production (n=14), melanin associated (n=33), titan cell-associated (n=6) and morphogenesis related (n=13) genes. Thus most of the genes involved in virulome analysis are multifaceted either directly or indirectly involved in more than one virulence trait. The list of genes in the study strain Cn belonging to each category is briefed in Figure 6.

In the virulome comparative study, H99 has the highest number of virulence genes with 99.26% (n=270) and lowest in Bt133 with 82.14% (n=224). However, both are clinical isolates. The Cn strain has 97.06% (n=264) of virulence genes. Though being an environmental isolate, VNII has 83.45% (n=227) virulence genes. For the comparative virulome analysis, the major VRGs polysaccharide capsule (*CAP10*, *CAP59*, *CAP60* and *CAP64*), melanin (*LAC1* and *LAC2*), thermotolerance (Hsp90 and *STU1*), cell adhesion (*ALG3*), iron acquisition (*CFT1*, *SIT1*, *CIR1* and RIM101), oxidative stress tolerance (*SOD1* and *SOD2*) and enzymes (*PLB1*, *URE1*, *YME1*, *RIM13*, *APR1*, *PIM1*, *SPP1*, *RBD1* and *APP2*) were used. The comparison revealed that these genes were found in almost all the *C. neoformans* sp. It was found that the

strains Cn, H99 and JEC21 have all the 24 major VRGs and could be more virulent than other strains (figure 7; table 1).

3.6. Resistome and Comparative Resistome Analysis

The AFR genes that are retrieved from the literature are examined for their presence in the study genomes using BLAST. The output significantly predicted AFR genes with a minimum of 99% identity and sequence coverage. Around 78 genes were predicted and observed as factors promoting antifungal resistance. The Cn has ABC transporters (n=6 genes), Major Facilitator Superfamily (n=13), drug target genes (n=12), Inherent resistance genes (n=2), Stress regulator genes (n=2), DNA mismatch genes (n=5) and transcriptional regulator genes (n=37) contributing to AFR (figure 8).

Among the 12 strains, the Cn was found to harbor a high number of AFR genes across certain classes of antifungals. Resistance to azole drugs were contributed by *ERG11, ERG24, ERG1, SRE1, AFR1, AFR2, PMR5, TUB1, SEY1, YOP1, TUB1* and *YAP1* genes. Resistance to echinocandins was contributed by *FKS1, CRM1,* and *Hsp90* genes. The gene for P4 ATPase and *CRM1* confers intrinsic resistance for echinocandins and the resistance to pyrimidine analog was conferred by the genes *FCY2*, *TMP1* and *FUR1.* Resistance to allylamines was provided by the gene *ERG1.* Conveying that the strain Cn (n=78; 100%) has more AFR genes than other isolates. The possible reason would be that the hybrid genome has been identified to have more AFR genes than other strains. Despite being an environmental strain the VNII has 53 AFR genes (68.83%).

Among 20 major antifungal resistance genes, 13 genes were commonly shared (100%) in all the 12 study strains. The genes *ERG11*, *ERG24*, *SEY1* and *ERG1* belonging to sterol biosynthesis provide resistance by the alteration of target protein structure, overexpression of target protein and increased activation of efflux pump. The gene for P4- ATPase and the gene *CRM1* provide inherent resistance by regulating the intracellular calcium concentration. The gene *FKS1* (β- glucan synthase; Cell wall synthesis) and Hsp90 (Heat shock protein 90; Cell wall regulation) provide resistance by alteration, under expression of the target protein and stress regulation. The genes *TMP1* and *FUR1* (Thymidylate synthase 1 and Uracil phosphoribosyltransferase 1; nucleotide biosynthesis) provide resistance by alteration and overexpression of target proteins. The gene *PMR5* (ABC transporter PMR5; ABC transporter) is the only common transporter protein that provides resistance to antifungal drugs (figure 9).

Among the 78 AFR genes, SNPs were detected in *ERG11* (lanosterol 14-α demethylase), *FKS1* (β- glucan synthase), *UXS1* (UDP- glucuronate decarboxylase 1), *FUR1* (Uracil phosphoribosyltransferase), *TUB1* (β tubulin) and *FCY2* (Cytosine permease) genes. Among them, *ERG11* was noted to have the highest number of mutations (n=9), followed by *UXS1* (n=2), *FKS1* (n=1), *FUR1* (n=1) and *FCY2* (n=1). The presence of SNPs in the 12 study strains is depicted in Figure 10; Table 2. In the *FKS1* gene, serine at the 643rd position was mutated to proline. This resulted in the conformational change in the lipid enriched molecules, as its role is to anchor the lipid molecules, and amino acid serine at 643rd position is involved in the binding site. In the *ERG11* gene phenylalanine (F) at 126th position has been changed to serine (S) and glycine at 464th position has been mutated to valine. Both these residues were not conferred to active site binding pocket. In the *TUB1*, histidine at the 6th position was changed to glutamine, glutamic acid at 98th position was mutated to glycine and phenylalanine at 200th position was mutated to asparagine. Glycine at the 190th position was mutated to glutamic acid in the ABC transporter gene *FUR1*. In all the 12 study genomes, though the mutation was observed, none of them were conferred to the binding pocket which is also concurrent with our *in vitro* susceptibility test where the study clinical strain showed sensitivity to fluconazole and amphotericin B (Rathore et al., 2017).

3.7. Pan-genome analysis

The pan-genome analysis aids in understanding the genome architecture of a species. The preliminary orthologous gene analysis showed the number of gene clusters, proteins and singleton genes present in each genome (figure 11). The overall read statistics of the orthologous analysis are provided in Table 3. Secondly, the upset chart of orthologous analysis represents the intersections of orthologous gene clusters, overlapping clusters and proteins between 12 study strains. From the upset chart of the gene cluster and overlapping cluster, it was known that there were about 4676 clusters shared commonly among the 12 genomes. The unique proteins and clusters (n=1 to n=89) defined in the upset chart (cluster count and protein) of each genome are given in Table 4. The occurrence chart visualizes the presence and absence of orthologous gene clusters and protein count across the compared species or datasets. The occurrence chart gives a correlation between the gene cluster and protein count. From the occurrence chart of the gene cluster, it was depicted that there were about n=4676 clusters (73.6%) containing n=58795 proteins (79.3%) present in all the study strains (figure 12, 13, 14). It also correlates the unique gene cluster of each genome with the protein count that is mentioned in Table 5. There were about 4304 single-copy clusters (single-copy orthologous

clusters) across the 12 genomes. The single copy clusters were categorized into biological process-associated (96.68%), molecular function-associated (2.23%) and cellular component-associated (1.08%) gene clusters (Figure 15; Table 6). The phylogenetic tree was constructed by the maximum likelihood method using the orthologous genes (figure 16). The midpoint was divided into two clades where the strain JEC21 (*C. neoformans var. neoformans*) was in clade1 and the rest in clade2. In clade 2 the strain VNII (environmental strain) was outgrouped and other clinical strains were clustered in 2 different clades. The phylogenetic tree constructed using the MLST sequence and orthologous sequence were merely similar. The gene expansion (gene duplication) and contraction (gene deletion) studies showed that there was no major expansion or contraction of genes. A pairwise heat map was generated based on the similarity of orthologous genes which shows that the strain Cn (clinical strain; sub sp. *grubii*) and VNII (environmental strain) have high similarity with JEC21 (clinical strain; sub sp. *neoformans*). The strain Bt65 (clinical strain) is similar to Bt 81 (clinical strain) and Bt89 (clinical strain) is similar to Bt133 (clinical strain) (figure 17). The result of CAFE5 also depicts there was no notable expansion or contraction in the genome of 12 study strains (Sun et al., 2023).

4. Discussion

In this study, we sequenced and assembled a genome of *C. neoformans* (Cn) collected from the microbiology lab of Trichy Medical College, Tamil Nadu, India sourced from the cryptococcal meningitis patient. Together with the already available 11 genomes of the species, we compared all 12 genomes and deciphered the phylogenics, resistome, virulome and pangenome of these cryptococcal strains. The genome size of all the study strains is 18.9 to 19.1 Mb and the GC contents ranged between 48 to 48.5%.

The ISHAM ITS barcoding helped to confirm the subspecies of the study strain Cn to be *grubii* (100% identity) which is in correlation with the constructed phylogenetic tree. The Cn study strain's subspecies identification also confirms that it is a member of the serotype A. Due to the lack of 100% matching sequences in the ISHAM database, we anticipate that the strain Cn might belong to a new ST type.

Our manual method of retrieving virulence genes from literature and BLAST analysis revealed the presence of 99.28% of virulence genes in H99. In various reported studies virulent nature of H99 was well demonstrated using animal models (Bosch et al., 2024; Fernandes et al., 2022). In addition, our recent demonstration of the disseminated cryptococcosis mice model also revealed the high virulent potency of H99. Confirming our manual method as a suitable one for virulome analysis, however, the creation of a virulome database is the need of the hour.

The virulome analysis depicts the presence of stress tolerance genes, which are pleiotropic and contribute to virulence regulation. The ability of *Cryptococcus* sp., to adapt to the host environment depends on how well it can maintain metabolic homeostasis under stress. The upregulation of carbohydrate transport, lipid transport, and autophagy, indicates that there is a significant need for the uptake of nutrients and energy production after host infection and phagocytosis. The *Cryptococcus* sp. has a set of autophagy genes that are essential for sustaining cellular integrity, homeostasis and survival (stress tolerance). Some of the metabolic pathways such as the pentose phosphate pathway which has been regulated by ATP were also upregulated under nitrosative stress conditions. This serves as an example that the metabolic pathways were likely to be induced by stress conditions. A recent study demonstrated that the transcription factor CIR1 (master iron regulator) and RIM101 integrate iron homeostasis with a multifarious of other functions including pH sensing, nutrient and stress signaling pathway, virulence factor elaboration and cell wall biogenesis (Ding et al., 2018; Jiang et al., 2020; Zhao et al., 2019).

All the 12 isolates showed most of the major VRGs to be found in their genome. The study strains C23, bt89 and Bt133 were found to lack the major virulence factor URE1, which serves to breech the BBB and host immune system. Except Cn, H99 and JEC21 all the other strains lack the gene LAC1 that contributes majorly to the melanin production, biofilm production and inhibition of phagocytic uptake. This is also concurrent with our previous study, in which the Cn was found to have a dominant capsule, melanin production and biofilm formation.

While comparing the virulome of the study strain Cn and the reference genome (H99), *ATG12*, *ATG16, CHS4*, Alkaline phosphatase, *UGD1*, *APP1*, *BOS1* and *RAS2* were found to be absent in the study strain which contributes to immune evasion. While comparing the virulome of the reference strain and the environmental strain, *ATG4*, *ATG9*, *ATG12*, *ATG14*, *ATG18*, *ATG26*, *CBP1*, *XPT1*, *CHS4*, *CHS6*, *CHS8*, *CDA3*, *URE1*, phosphatidic acid phosphatase, *SIP4*, *CAN1*, *FRR1*, *SHO1*, *CLN1*, *SEC61B*, *SEC16*, *LAC1*, copper transport protein 86, potassium transport protein, *PDK1*, Outer cell wall protein RRT8, Yeast cell wall synthesis Kre9/Knh1-like N-terminal domain-containing protein, TatD DNase family Scn1, *CAN2*, homoserine kinase, Homoserine O-acetyltransferase, α-amylase, Alkaline phosphatase, 2-Methylcitrate synthase, *UGD1*, RNaseIII, Inositolphosphorylceramide synthase subunit Kei1-domain-containing protein, *BIM1*, *VMA21*, *UGT1*, *BOS1*, *RUB1*, *SCH9*, *RAS2* were found to be absent in the environmental strain that has protease, urease, laccase, autophagy-related and cellular function.

Even though all the strains contained the AFR genes with certain mutations, most of the mutational sites do not belong to the active site binding pocket. As additional supporting data, the *in vitro* investigation on the study strain and the reference strain (H99) demonstrated that both the strains were susceptible to amphotericin B (polyenes) and fluconazole (azole). With this knowledge, it was anticipated that a mutation that occurs outside the active site binding pocket might not be the cause of AFR. This comparative genomic study of *C. neoformans* on virulence and AFR gives insights into the potent targets for the development of novel antifungal drugs.

The pan-genome analysis of the *Cryptococcus neoformans* gives various insights into the orthologous gene and protein levels. The number of gene clusters and proteins are not necessarily correlated. Though the strain KBCN0142 has the highest number of cluster (n=5995), the strain Bt81 has the highest number of proteins (n=6366). This might be due to the fact of high redundancy, high similarity among the proteins, stringent clustering parameters, and the presence of paralogous genes. The presence of singleton proteins in the genomes shows that they are essential proteins that appear unique within the dataset and don't have any closely related homologs. The strain Cn has the highest number of singletons up to 136. The singleton showcases the unique protein function and evolutionary significance. The orthologous analysis performed demonstrated the intersection of gene clusters and proteins; unique clusters and the proteins of each genome. Though there were unique clusters and unique proteins present in all the 12 study strain genomes, as represented by the upset charts of cluster count and protein, the occurrence chart where the gene clusters are correlated with the protein shows the absence of a unique cluster or protein in H99, C23, KBCN0140, KBCN0142 and SACN00B1. The single-copy clusters depict the group of orthologous genes that contains exactly one copy of a gene in the cluster. Such clusters are exceptionally important in comparative genomics and evolutionary study as they represent the conserved genes that reduce the noise from gene duplication or loss events in the compared species. The single-copy gene clusters were categorized based on functional nature as a biological process, molecular process and cellular component. Among them, the biological process-related GO cluster contributes to a maximum of 96.69% (n=17262), molecular function contributes to about 2.23% (n=398) and finally, the cellular component contributes to about 1.08% (n=193). This shows that the biological process-related genes are more conserved across the species. Though the strain JEC21 belongs to the subspecies *neoformans* and the strain Cn belongs to sub sp. *grubii* the pairwise heat map shows high similarity between them which might be due to the fact that both the subspecies were

descendents of the same ancestor. The phylogenetic tree constructed with the orthologous genes was considered to show more evolutionary relatedness than the tree constructed with the housekeeping genes. The reason behind the statement is the comprehensive representation of the genome, robustness to homoplasy, reduced noise of gene duplication or gene loss and improved alignment quality. These factors reflect the true evolutionary relationships among the species. In this study, the phylogenetic tree constructed using the MLST genes and orthologous genes are merely similar. The gene expansion and contraction studies showed 0+/0- for all the strains, showing that there is no unique expansion (additional gene copies have been gained in their lineage) or contraction (number of gene copies lost in the lineage). The possibility for this lack of expansion/contraction might be due to the presence of stable gene families (the gene families in the species are stable concerning other species showing no significant changes in size due to duplication or loss events) or equal evolutionary changes (either the expansion/contraction occurring uniformly/ not at all occurring across all the species being analyzed leading to unique changes in a single lineage)

5. Conclusion

In conclusion, this was the first study to report the complete genome sequence of clinical *Cryptococcus* sp. from India and to study comparative genomics of strains from different geographical locations worldwide. All these isolates showed similar patterns of virulence but differed in the levels of melanization, stress, evasion of BBB and pathogenicity. The pathogenicity assessment reveals that the study strains Cn and H99 are more pathogenic than the other study strains. This study provides the first comparative genome analysis of *C. neoformans* based on virulence, AFR and pan-genome. These genomic resources could be further explored through molecular, functional and evolutionary analysis to understand the virulence and treat cryptococcal infections efficiently. The study advocates the need for bioinformatic tools for resistome and virulome analysis.

Acknowledgment

We thank SASTRA Deemed University for providing the research facilities and infrastructure.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The author(s) reported there is no funding associated with the work featured in this article.

References

Alastruey-Izquierdo, A., Martín-Galiano, A.J., 2023. The challenges of the genome-based identification of antifungal resistance in the clinical routine. Front. Microbiol. 14, 1–8. https://doi.org/10.3389/fmicb.2023.1134755

Almeida, F., Wolf, J.M., 2015. Virulence-Associated Enzymes of. https://doi.org/10.1128/EC.00103-15.Address

Arastehfar, A., Gabaldón, T., Garcia-Rubio, R., Jenks, J.D., Hoenigl, M., Salzer, H.J.F., Ilkit, M., Lass-Flörl, C., Perlin, D.S., 2020. Drug-resistant fungi: An emerging challenge threatening our limited antifungal armamentarium. Antibiotics 9, 1–29. https://doi.org/10.3390/antibiotics9120877

Azevedo, R., Rizzo, J., Rodrigues, M., 2016. Virulence Factors as Targets for Anticryptococcal Therapy. J. Fungi 2, 29. https://doi.org/10.3390/jof2040029

Bosch, C., Toplis, B., Van Staden, A.D.P., Volschenk, H., Smith, C., Dicks, L., Botha, A., 2024. Exposure of Cryptococcus neoformans to low nitrogen levels enhances virulence. Int. Microbiol. https://doi.org/10.1007/s10123-024-00504-y

Brandão, F., Esher, S.K., Ost, K.S., Pianalto, K., Nichols, C.B., Fernandes, L., Bocca, A.L., Poças-Fonseca, M.J., Alspaugh, J.A., 2018. HDAC genes play distinct and redundant roles in Cryptococcus neoformans virulence. Sci. Rep. 8, 1–17. https://doi.org/10.1038/s41598-018-21965-y

Brown, S.M., Campbell, L.T., Lodge, J.K., 2007. Cryptococcus neoformans, a fungus under stress. Curr. Opin. Microbiol. 10, 320–325. https://doi.org/10.1016/j.mib.2007.05.014

Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND 12. https://doi.org/10.1038/nmeth.3176

Cao, C., Wang, Y., Husain, S., Soteropoulos, P., Xue, C., 2019. Echinocandin Resistance in Cryptococcus neoformans 1–18.

Casalini, G., Giacomelli, A., Antinori, S., 2024. The WHO fungal priority pathogens list: a crucial reappraisal to review the prioritisation. The Lancet Microbe 5, 717–724. https://doi.org/10.1016/S2666-5247(24)00042-9

Chayakulkeeree, M., Perfect, J.R., 2006. Cryptococcosis. Infect. Dis. Clin. North Am. 20,

507–544. https://doi.org/10.1016/j.idc.2006.07.001

Chen, Y., Shi, Z.W., Strickland, A.B., Shi, M., 2022. Cryptococcus neoformans Infection in the Central Nervous System : The Battle between Host and Pathogen.

Coelho, C., Bocca, A.L., Casadevall, A., 2014. The Tools for Virulence of Cryptococcus neoformans, Advances in Applied Microbiology. https://doi.org/10.1016/B978-0-12-800261-2.00001-3

Ding, H., Caza, M., Dong, Y., Arif, A.A., Horianopoulos, L.C., Hu, G., Johnson, P., Kronstad, J.W., 2018. ATG genes influence the virulence of Cryptococcus neoformans through contributions beyond core autophagy functions. Infect. Immun. 86, 1–24. https://doi.org/10.1128/IAI.00069-18

Diniz-lima, I., Marques, L., Silva-junior, E.B., Cristina, J., Freire-de-lima, L., Nascimento, D.O., Morrot, A., Previato, J.O., Mendonça-previato, L., Decote-ricardo, D., Freire-de-lima, C.G., 2022. applied sciences Cryptococcus : History , Epidemiology and Immune Evasion.

do Carmo, F.N., de Camargo Fenley, J., Garcia, M.T., Rossoni, R.D., Junqueira, J.C., de Barros, P.P., Scorzoni, L., 2022. Cryptococcus spp. and Cryptococcosis: focusing on the infection in Brazil. Brazilian J. Microbiol. 1321–1337. https://doi.org/10.1007/s42770-022-00744-y

Fernandes, K.E., Fraser, J.A., Carter, D.A., 2022. Lineages Derived from Cryptococcus neoformans Type Strain H99 Support a Link between the Capacity to Be Pleomorphic and Virulence. MBio 13, 1–15. https://doi.org/10.1128/mbio.00283-22

Firacative, C., Meyer, W., Castañeda, E., 2021. Cryptococcus neoformans and Cryptococcus gattii species complexes in latin america: A map of molecular types, genotypic diversity, and antifungal susceptibility as reported by the latin american cryptococcal study group. J. Fungi 7. https://doi.org/10.3390/jof7040282

Gow, N.A.R., Johnson, C., Berman, J., Coste, A.T., Cuomo, C.A., Perlin, D.S., Bicanic, T., Harrison, T.S., Wiederhold, N., Bromley, M., Chiller, T., 2022. The importance of antimicrobial resistance in medical mycology 1–12. https://doi.org/10.1038/s41467-022-32249-5

Ha, S.Y.S., Lim, J., Kwon, S., Chun, J., 2017. A large-scale evaluation of algorithms to

calculate average nucleotide identity. Antonie Van Leeuwenhoek 110, 1281–1286. https://doi.org/10.1007/s10482-017-0844-4

Hoff, K.J., Lomsadze, A., Stanke, M., Borodovsky, M., 2021. BRAKER2 : automatic eukaryotic genome annotation with GeneMark-EP + and AUGUSTUS supported by a ´ s ˇ Br una Tom a 3, 1–11. https://doi.org/10.1093/nargab/lqaa108

Jiang, S.T., Chang, A.N., Han, L.T., Guo, J.S., Li, Y.H., Liu, T.B., 2020. Autophagy Regulates Fungal Virulence and Sexual Reproduction in Cryptococcus neoformans. Front. Cell Dev. Biol. 8. https://doi.org/10.3389/fcell.2020.00374

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018. MEGA X : Molecular Evolutionary Genetics Analysis across Computing Platforms. https://doi.org/10.1093/molbev/msy096

Kwon-chung, K.J., Fraser, J.A., Doering, T.L., Wang, Z.A., Janbon, G., Idnurm, A., Bahn, Y., 2014. gattii , the Etiologic Agents of Cryptococcosis. Cold Spring Harb. Prespectives Med. 4, a019760.

Letunic, I., Bork, P., 2024. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res. 78–82. https://doi.org/10.1093/nar/gkae268

Lu, T., Yao, B., Zhang, C., 2012. Original article DFVF : database of fungal virulence factors 2012, 1–4. https://doi.org/10.1093/database/bas032

Malachowski, A.N., Yosri, M., Park, G., Bahn, Y.S., He, Y., Olszewski, M.A., 2016. Systemic approach to virulence gene network analysis for gaining new insight into cryptococcal virulence. Front. Microbiol. 7, 1–14. https://doi.org/10.3389/fmicb.2016.01652

Manni, M., Berkeley, M.R., Seppey, M., Zdobnov, E.M., 2022. BUSCO : Assessing Genomic Data Quality and Beyond 1–41. https://doi.org/10.1002/cpz1.323

Mau, B., Perna, N.T., 2014. arXiv : 0910 . 5780v1 [q-bio . GN] 30 Oct 2009 Progressive Mauve : Multiple alignment of genomes with gene flux and rearrangement.

Maziarz, E.K., Perfect, J.R., 2016. Cryptococcosis. Infect. Dis. Clin. North Am. 30, 179–206. https://doi.org/10.1016/j.idc.2015.10.006

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., Kanehisa, M., 2007. KAAS: An automatic

genome annotation and pathway reconstruction server. Nucleic Acids Res. 35, 182–185. https://doi.org/10.1093/nar/gkm321

Park, Y.D., Sun, W., Salas, A., Antia, A., Carvajal, C., Wang, A., Xu, X., Meng, Z., Zhou, M., Tawa, G.J., Dehdashti, J., Zheng, W., Henderson, C.M., Zelazny, A.M., Williamson, P.R., 2016. Identification of multiple cryptococcal fungicidal drug targets by combined gene dosing and drug affinity responsive target stability screening. MBio 7. https://doi.org/10.1128/mBio.01073-16

Patiño, L.H., Muñoz, M., Ramírez, A.L., Vélez, N., Escandón, P., Parra-Giraldo, C.M., Ramírez, J.D., 2023. A Landscape of the Genomic Structure of Cryptococcus neoformans in Colombian Isolates. J. Fungi 9. https://doi.org/10.3390/jof9020135

QIAseq ® FX DNA Library Kit Handbook, 2023.

Rajasingham, R., Govender, N.P., Jordan, A., Loyse, A., Shroufi, A., Denning, D.W., Meya, D.B., Chiller, T.M., 2022. Articles The global burden of HIV-associated cryptococcal infection in adults in 2020 : a modelling analysis. Lancet Infect. Dis. 3099, 1–8. https://doi.org/10.1016/S1473-3099(22)00499-6

Rathore, S.S., Isravel, M., Vellaisamy, S., Raj, D., 2017. Exploration of Antifungal and Immunomodulatory Potentials of a Furanone Derivative to Rescue Disseminated Cryptococosis in Mice. Sci. Rep. 1–14. https://doi.org/10.1038/s41598-017-15500-8

Rathore, S.S., Sathiyamoorthy, J., Lalitha, C., Ramakrishnan, J., 2022. A holistic review on Cryptococcus neoformans. Microb. Pathog. 166, 105521. https://doi.org/10.1016/j.micpath.2022.105521

Sabiiti, W., May, R.C., 2012. Mechanisms of infection by the human fungal pathogen. Future Microbiol. 7, 1297–1313.

Snelders, E., Moyrand, F., Sturny-Leclère, A., Vernel-Pauillac, F., Volant, S., Janbon, G., Alanio, A., 2022. The role of glycosylphosphatidylinositol (gpi) anchored proteins in Cryptococcus neoformans. Microbes Infect. 24. https://doi.org/10.1016/j.micinf.2022.105016

Stanke, M., Morgenstern, B., 2005. AUGUSTUS : a web server for gene prediction in eukaryotes that allows user-defined constraints 33, 465–467. https://doi.org/10.1093/nar/gki458

Sun, J., Lu, F., Luo, Y., Bie, L., Xu, L., Wang, Y., 2023. OrthoVenn3: An integrated platform for exploring and visualizing orthologous data across genomes. Nucleic Acids Res. 51, W397–W403. https://doi.org/10.1093/nar/gkad313

Verma, S., Shakya, V.P.S., Idnurm, A., 2018. Exploring and exploiting the connection between mitochondria and the virulence of human pathogenic fungi. Virulence 9, 426–446. https://doi.org/10.1080/21505594.2017.1414133

WHO, 2021. Indian Priority Pathogen List 1–24.

Xess, I., Pandey, M., Dabas, Y., Agarwal, R., Das, S., Srivastava, P.M.V., Thakur, R., Sharma, S., Mani, P., Biswas, A., Bhowmik, D., Das, B.K., Singh, R., Panda, P., Singh, G., Phukan, A.C., Yhome, M., Valsan, C., Kindo, A.J., 2021. Multilocus Sequence Typing of Clinical Isolates of Cryptococcus from India. Mycopathologia 186, 199–211. https://doi.org/10.1007/s11046-020-00500-6

Xu, G., Xu, T., Zhu, R., Wang, H., Li, J., 2018. LR Gapcloser : a tiling path-based gap closer that uses long reads to complete genome assembly 1–14. https://doi.org/10.1093/gigascience/giy157

Zaragoza, O., 2019. Basic principles of the virulence of Cryptococcus. Virulence 10, 490–501. https://doi.org/10.1080/21505594.2019.1614383

Zhang, S., Hacham, M., Panepinto, J., Hu, G., Shin, S., Zhu, X., Williamson, P.R., 2006. The Hsp70 member, Ssa1, acts as a DNA-binding transcriptional co-activator of laccase in Cryptococcus neoformans. Mol. Microbiol. 62, 1090–1101. https://doi.org/10.1111/j.1365-2958.2006.05422.x

Zhao, X., Feng, W., Zhu, X., Li, C., Ma, X., Li, X., 2019. Conserved Autophagy Pathway Contributes to Stress Tolerance and Virulence and Differentially Controls Autophagic Flux Upon Nutrient Starvation in Cryptococcus neoformans 10, 1–14. https://doi.org/10.3389/fmicb.2019.02690

Zimin, A. V, Marc, G., Puiu, D., Roberts, M., Salzberg, S.L., Yorke, J.A., 2013. Genome analysis The MaSuRCA genome assembler 29, 2669–2677. https://doi.org/10.1093/bioinformatics/btt476

List of Figures

Figure 1: A flow chart on the comprehensive methodology used in the study

Figure 2: GO annotation by DIAMOND database: the Gene ontology of the study strain Cn

Figure 3: Pathway analysis by KAAS database: Top 10 pathway functions of the study strain Cn

Figure 4: Genome identity search: (a) ANI Calculation of the study strain Cn and reference strain H99; (b) Dot plot comparison using D GENIES and (c) Linear genome image using MAUVE

Figure 5: Phylogenetic tree based on SNP: SNP phylogeny was constructed by the maximum likelihood method with MegaX. 1000 bootstrap replicates were calculated to assess robustness. The midpoint of the phylogenetic tree was divided into two major clades and further delineated into many clusters of *C. neoformans*

Figure 6: VRGs present in the study strain Cn: The chart represents the (n=264) VRGs of Cn that were categorized into stress tolerance, cell wall-associated, cell membrane-associated, capsule-associated, morphogenesis associated, cell adhesion, immune evasion, iron acquisition, protease, nutrient acquisition, copper homeostasis, host cell membrane disruption and melanin associated genes.

Figure 7: Comparative virulome analysis: The graph represents the comparative study of the major VRGs present in the (n=12) study strains. The colored region represents the presence of genes.

Figure 8: AFR genes present in the study strain Cn: The chart represents the (n=78) AFRs of the Cn that were categorized into transporter, structural, DNA mismatch, nucleic acid synthesis, stress response and transcription factors.

Figure 9: Comparative resistome analysis: The graph represents the comparative study of the major AFR genes present in the (n=12) study strains. The colored region represents the presence of genes.

Figure 10: Comparative analysis of SNPs in the AFR genes: The graph represents the comparative study of the SNPs present in the AFR genes of the (n=12) study strains. The colored region represents the presence of genes.

Figure 11: The preliminary orthologous analysis that shows the number of gene clusters (a), proteins (b) and singletons (c) present in each study strain

Figure 12: The pictorial representation of the occurrence chart that depicts the presence of orthologous gene clusters present and shared among the 12 genomes

Figure 13: The pictorial representation of the occurrence chart that depicts the presence of overlapping gene clusters present and shared among the 12 genomes.

Figure 14: The pictorial representation of the occurrence chart that depicts the presence of proteins present and shared among the 12 genomes.

Figure 15: Single-copy gene clusters present in 12 genomes classified as cellular component genes (a), biological process gene (b) and molecular function genes (c)

Figure 16: Phylogenetic tree based on orthologous genes: Orthologous gene-based phylogeny was constructed by the maximum likelihood method with CAFE5. The midpoint of the phylogenetic tree was divided into two major clades and further delineated into many clusters of *C. neoformans.*

Figure 17: Pairwise heatmap: The intensity of the colour gradient shows the similarity between the species

List of supplementary figures:

Supplementary figure 1: Agarose gel electrophoresis of isolated cryptococcal DNA: A- represents the HIND III digested lambda ladder and B- represents the isolated DNA of Cn strain

Supplementary figure 2: Tape station profiling: Fragment size distribution analysis of the prepared library

List of tables:

Table 1: The major VRGs of *C. neoformans* study strains

Virulence genes	Protein Name	Functions
CAP 10	Capsular associated protein 10	capsule synthesis and maintanence
CAP 59	Capsular associated protein 59	capsule synthesis and maintanence, Biofilm production, GXM transport
CAP 60	Capsular associated protein 60	capsule synthesis and maintanence
CAP 64	Capsular associated protein 64	capsule synthesis and maintanence
LAC1	Laccase1	production of melanin; regulates biofilm formation; Inhibits host cell phagocytosis
LAC2	Laccase2	production of melanin; inhibits host cell phagocytosis
STU1	protein STU1	stress responses; morphogenesis; capsule regulation
Hsp90	heat shock protein 90	regulates heat shock, oxidative stress and various physiological/ environmental stress including stress of antifungal agents; regulates capsule formation, melanin production; Regulates host immune evasion
URE1	urease1	facilitates nitrogen source utilization; alkalinization of the environment; calcium carbonate production; production of biofilm
PLB1	phospholipase B1	Immune evasion; fungal cell adhesion to host cells; nutrient acquisition; regulates titan cell formation
YME1 (C361_06299	ATP dependent metalloprotease	Catalyzes mechanism that involves a metal and related to virulence (stress response)
RIM13 (C361_05602)	Zinc metalloprotease	Catalyzes mechanism that involves a metal and related to virulence (stress response)
APR1 (C361_03288)	Aspartyl protease	Involved in host immune evasion and food acquisition
PIM1	Protease	host immune evasion by proteolysis
SPP1 (CNAG_05742)	S2P endopeptidase	host immune evasion by proteolysis
RBD1 (CNAG_04056)	Serine peptidase	fungal cell adhesion to host cell; immune evasion; nutrient acquisition
APP2 (CNAG_04009)	Aminopeptidase 2	Proteolysis; nutrient acquisition; capsule formation; immune evasion
CFT1	CFT1 protein, Cryptococcus Fe Transporter (high affinity iron uptake)	iron acquisition
SIT1	Siderophore-iron transporter Str1	Iron acquisition
CIR1	GATA type zinc finger protein asd-4	regulates iron homeostasis, capsule formation, melanin production, stress response, and host immune evasion and pathogenesis

RIM101	pH-response transcription factor pacC/RIM101	regulates genes encoding cell wall-modifying enzymes such as chitinases, glucanases, and cell wall proteins; capsule production
ALG3	Dol-P-Man:Man(5)GlcNAc(2)-PP-Dol alpha-1,3-mannosyltransferase	Maintains cell wall integrity; Fungal cell adhesion to host cell; Immune evasion; Nutrient acquisition
SOD2	Mn superoxide dismutase	Oxidative stress tolerance (Catalyzes dismutation of toxic superoxide, converting superoxide to hydrogen peroxide and oxygen and related to virulence)
SOD1	Copper zinc superoxide dismutase	Oxidative stress tolerance (Catalyzes dismutation of toxic superoxide, converting superoxide to hydrogen peroxide and oxygen and related to virulence)

Table 2: The SNPs present in AFR genes of *C. neoformans* study strains

Genes	**Residue in protein**	**Mutation detected differently than proposed in journals**
ERG11(9)		
Y145F	active site residue	-
G484S	active site residue	-
S460T	not an active site residue	-
A457V	not an active site residue	-
I99V	not an active site residue	-
Y132F	active site residue	-
K143R	not an active site residue	-
F126L	not an active site residue	F126S
G464S	not an active site residue	G464V
FKS1 (1)		
S643P	not an active site residue	-
TUB1 (3)		
H6Q	not an active site residue	H6R
E98	not an active site residue	E98G
F200	not an active site residue	F200N
UXS1 (2)		
D306G	not an active site residue	-
Y217C	not an active site residue	-
FUR1 (3)		
G190D	not an active site residue	-
G87D	not an active site residue	-
G190E	not an active site residue	-
FCY2 (1)		
A157V	not an active site residue	-

Table 3: The overall read statistics of orthologous gene

All clusters	6337
Overlapping clusters	170
Single copy clusters	4304
All proteins	74110
All singletons	991
Percentage of singletons	1.34%

Table 4: The unique orthologous cluster and proteins present in the 12 genomes- Upset table where the unique clusters and protein count are independent

Strain	Unique cluster	Unique protein
Cn	9	9
H99	89	89
JEC21	19	19
Bt65	4	4
Bt81	2	2
Bt89	1	1
Bt133	1	1
C23	88	88
VNII	6	6
KBCN0140	87	87
KBCN0142	89	89
SACN00B1	87	87
Total	**482**	**482**

Table 5: The unique orthologous cluster and proteins present in the 12 genomes- Occurance table where the unique clusters are correlated with the protein count

S.no.	Strain name	Unique clusters	Unique proteins
1.	JEC21	19	108
2.	Cn	9	27
3.	Bt65	4	84
4.	Bt81	2	9
5.	VNII	6	17
6.	Bt133	1	2
7.	Bt89	1	2
Total		**42**	**249**

Table 6: The single copy number genes present across the 12 genomes.

(a) Biological process cluster

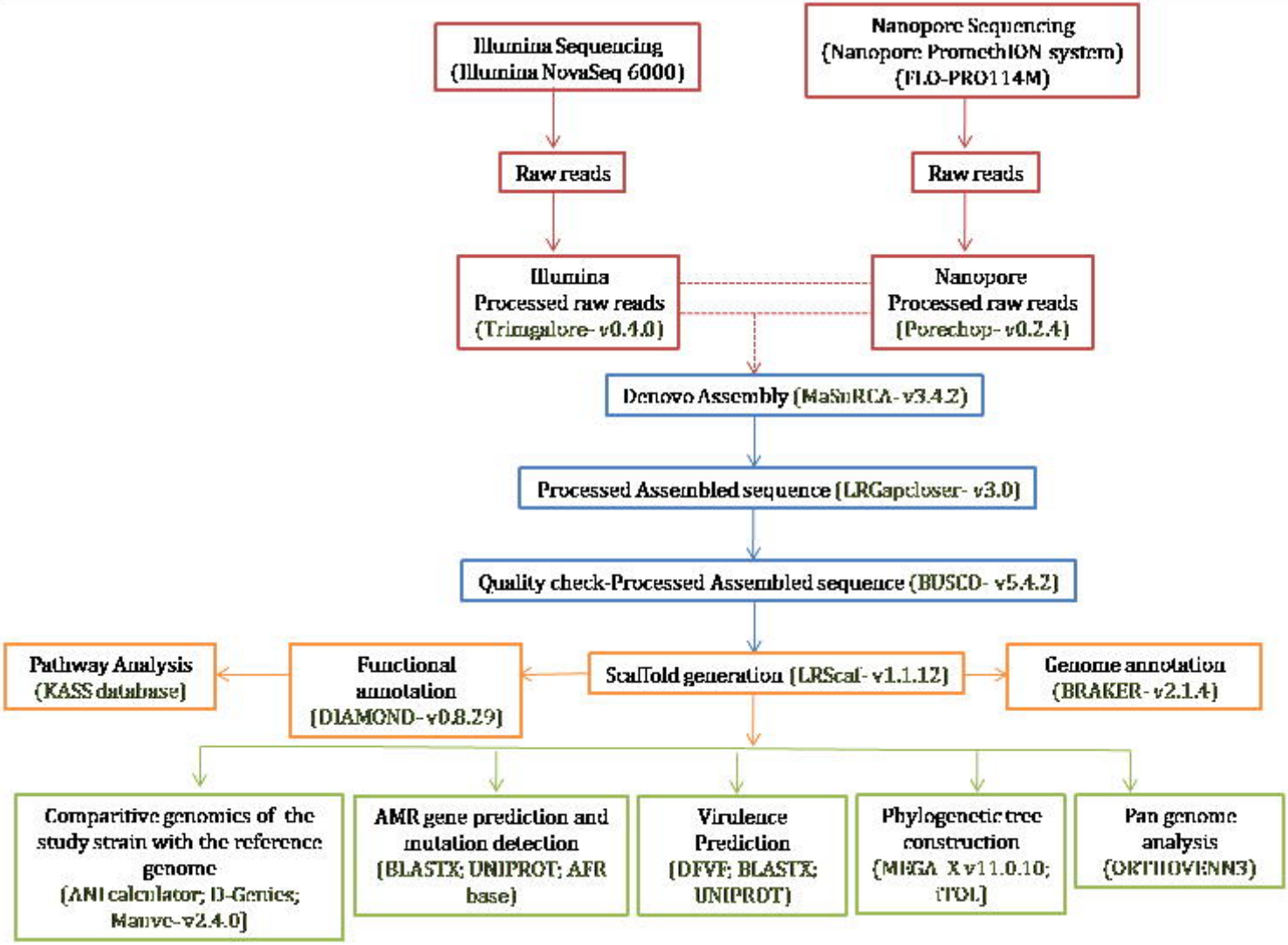
GO annotation number	**Count of unique input accession**	**Function**	**% contribution**
GO:0008150	1762	Biological process	18.48
GO:0008152	1576	metabolic process	16.53
GO:0044237	1114	Cellular metabolic process	11.68
GO:0009987	1037	Cellular process	10.87
GO:0043170	968	Macromolecule metabolic process	10.15
GO:0044238	783	Primary metabolic process	8.21
GO:0006807	778	Nitrogen compound metabolic process	8.16
GO:0016070	563	RNA metabolic process	5.90
GO:0065007	496	Biological regulation	5.20
GO:0046483	459	Heterocycle metabolic process	4.81
Total	**9536**	-	-

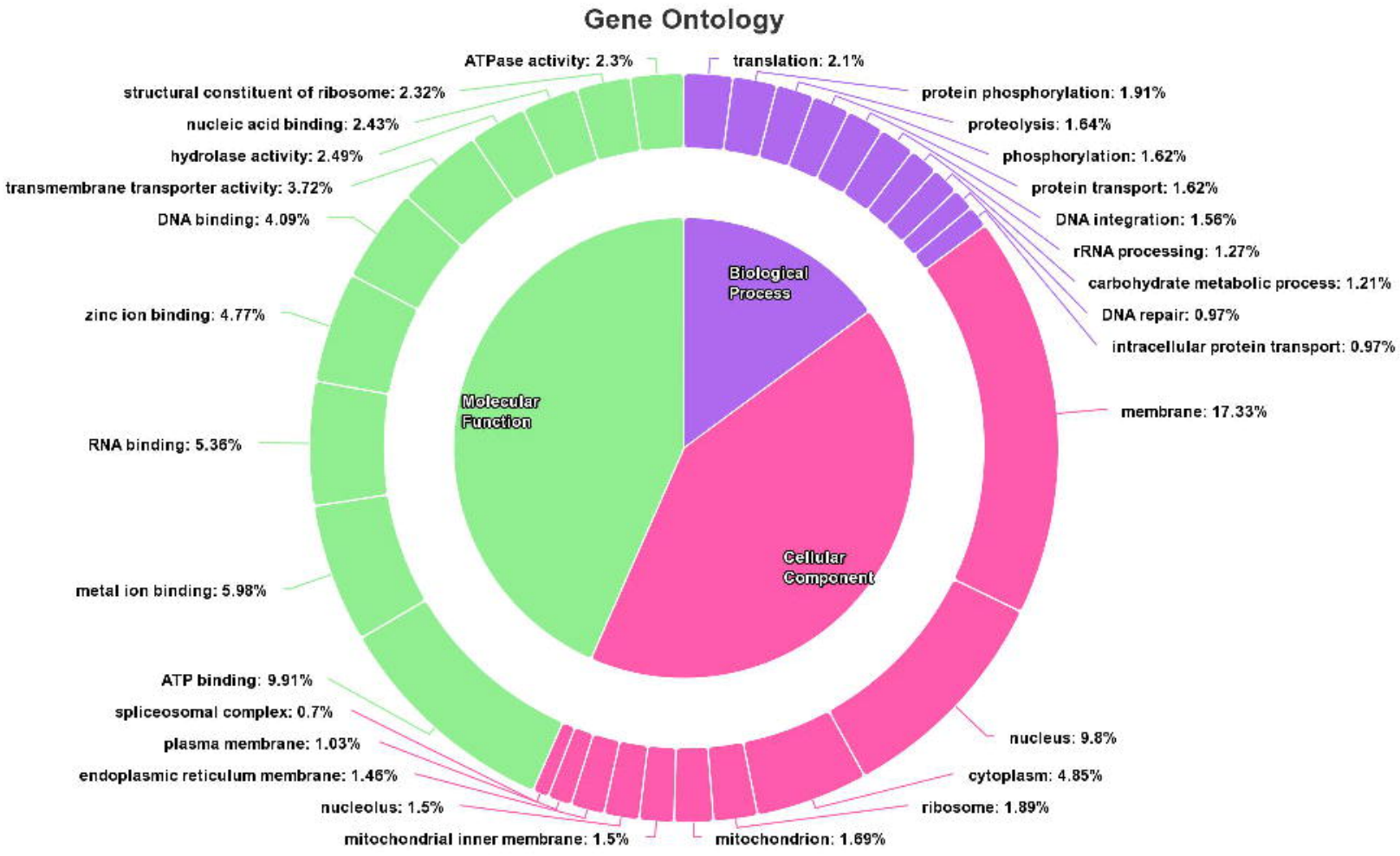
(6b) Molecular function cluster

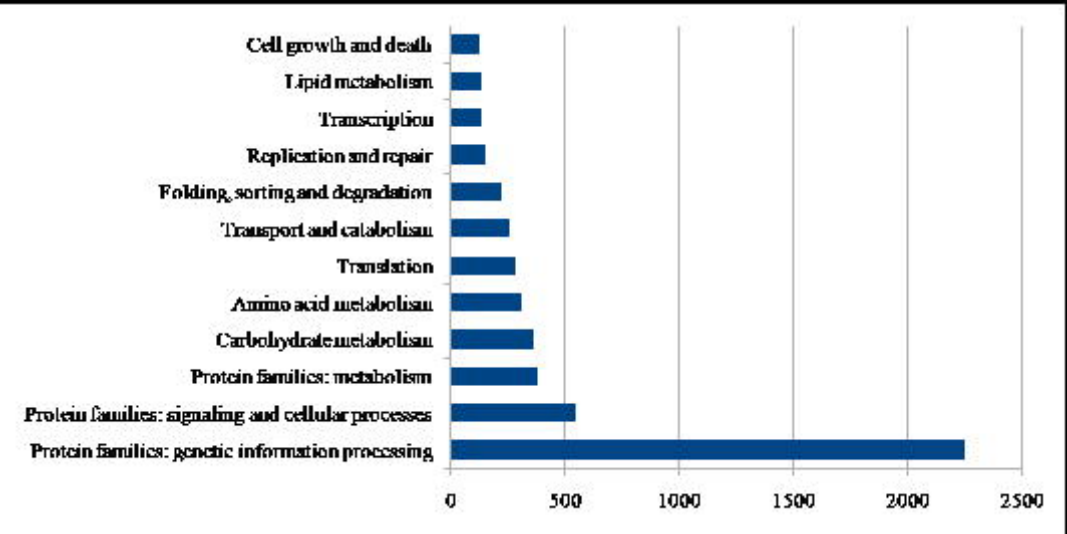
Molecular function			
GO annotation number	**Count of unique input accession**	**Function**	**% contribution**
GO:0016787	53	Hydrolase activity	16.11
GO:0016491	50	Oxidoreductase activity	15.20
GO:0016740	38	Transferase activity	11.55
GO:0043167	38	Ion binding	11.55
GO:0003676	35	Nucleic acid binding	10.64
GO:0003674	31	Molecular function	9.42
GO:0005215	26	Transporter activity	7.90
GO:0008233	26	Peptidase activity	7.90
GO:0000166	16	Nucleotide binding	4.86
GO:0048037	16	Cofactor binding	4.86
Total	**329**	-	-

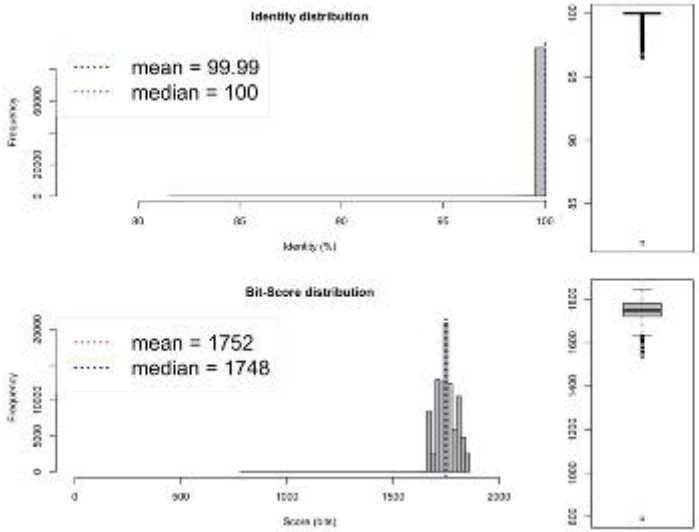
(6c) Cellular component cluster

Cellular component function			
GO annotation number	**Count of unique input accession**	**Function**	**% contribution**
GO:0016020	36	Membrane	19.56
GO:0005575	31	Cellular compartment	16.85
GO:0044464	27	Call part	14.67
GO:0005634	24	Nucleus	13.04
GO:0005622	19	Intracellular	10.33
GO:0043226	14	Organelle	7.61
GO:0043229	14	Intracellular organelle	7.61
GO:0005739	10	Mitochondrion	5.43
GO:0005576	5	Extracellular region	2.71
GO:0005794	4	Golgi apparatus	2.17
Total	**184**	-	-

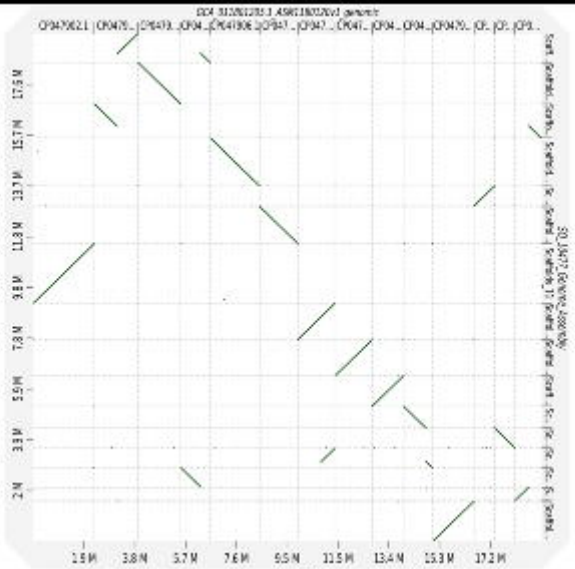

Illumina Sequencing
(Illumina NovaSeq 6000)
Nanopore Sequencing
(Nanopore PromethION system)
(FLO-PRO114M)
Raw reads
Raw reads
Illumina
Processed raw reads
(Trimgalore- v0.4.0)
Nanopore
Processed raw reads
(Porechop- v0.2.4)
Denovo Assembly (MaSuRCA- v3.4.2)
Processed Assembled sequence (LRGapcloser- v3.0)
Quality check-Processed Assembled sequence (BUSCO- v5.4.2)
Pathway Analysis
(KASS database)
Functional
annotation
(DIAMOND- v0.8.29)
ScaffOld generation (LRScaf- v1.1.12)
Genome annotation
(BRAKER- v2.1.4)
Comparitive genomics of the
study strain with the reference
genome
(ANI calculator; D-Genics;
Mauve- v2.4.0)
AMR gene prediction and
mutation detection
(BLASTX; UNIPROT; AFR
base)
Virulence
Prediction
(DFVF; BLASTX;
UNIPROT)
Phylogenetic tree
construction
(MEGA X v11.0.10;
iTOL)
Pan genome
analysis
(ORTHOVENN3)


Gene Ontology
Biological Process
Molecular Function
Cellular Component
translation: 2.1%
protein phosphorylation: 1.91%
proteolysis: 1.64%
phosphorylation: 1.62%
protein transport: 1.62%
DNA integration: 1.56%
rRNA processing: 1.27%
carbohydrate metabolic process: 1.21%
DNA repair: 0.97%
intracellular protein transport: 0.97%
membrane: 17.33%
nucleus: 9.8%
cytoplasm: 4.85%
ribosome: 1.89%
mitochondrion: 1.69%
mitochondrial inner membrane: 1.5%
nucleolus: 1.5%
endoplasmic reticulum membrane: 1.46%
plasma membrane: 1.03%
spliceosomal complex: 0.7%
ATP binding: 9.91%
metal ion binding: 5.98%
RNA binding: 5.36%
zinc ion binding: 4.77%
DNA binding: 4.09%
transmembrane transporter activity: 3.72%
hydrolase activity: 2.49%
nucleic acid binding: 2.43%
structural constituent of ribosome: 2.32%
ATPase activity: 2.3%
Highcharts.com

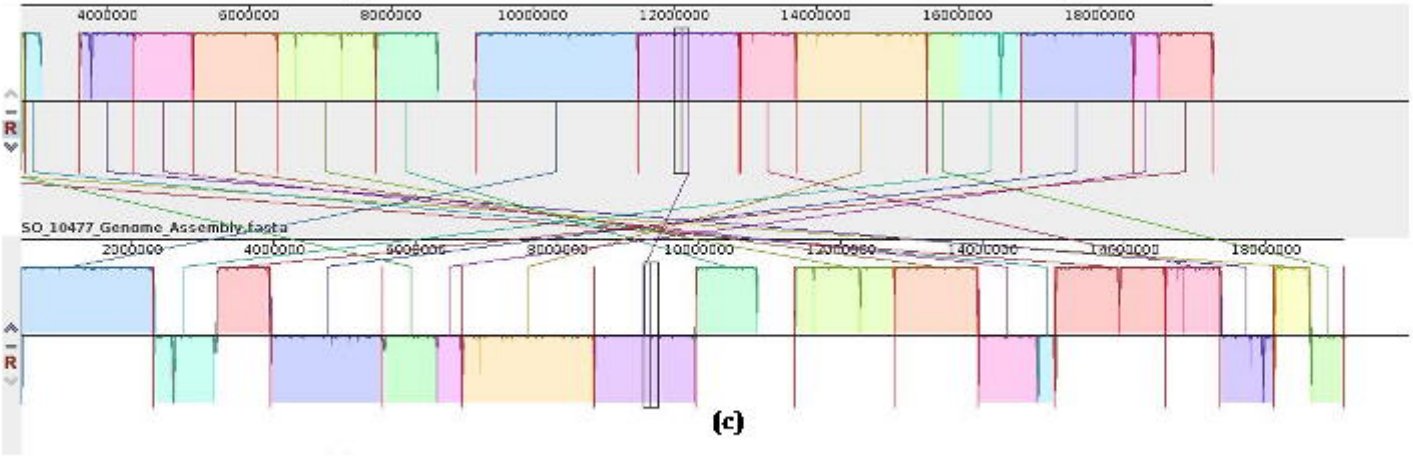

Cell growth and death
Lipid metabolism
Transcription
Replication and repair
Folding, sorting and degradation
Transport and catabolism
Translation
Amino acid metabolism
Carbohydrate metabolism
Protein families: metabolism
Protein families: signaling and cellular processes
Protein families: genetic information processing
0
500
1000
1500
2000
2500


Identity distribution
mean = 99.99
median = 100
Frequency
Identity (%)
Bit-Score distribution
mean = 1752
median = 1748
Frequency
Score (bits)

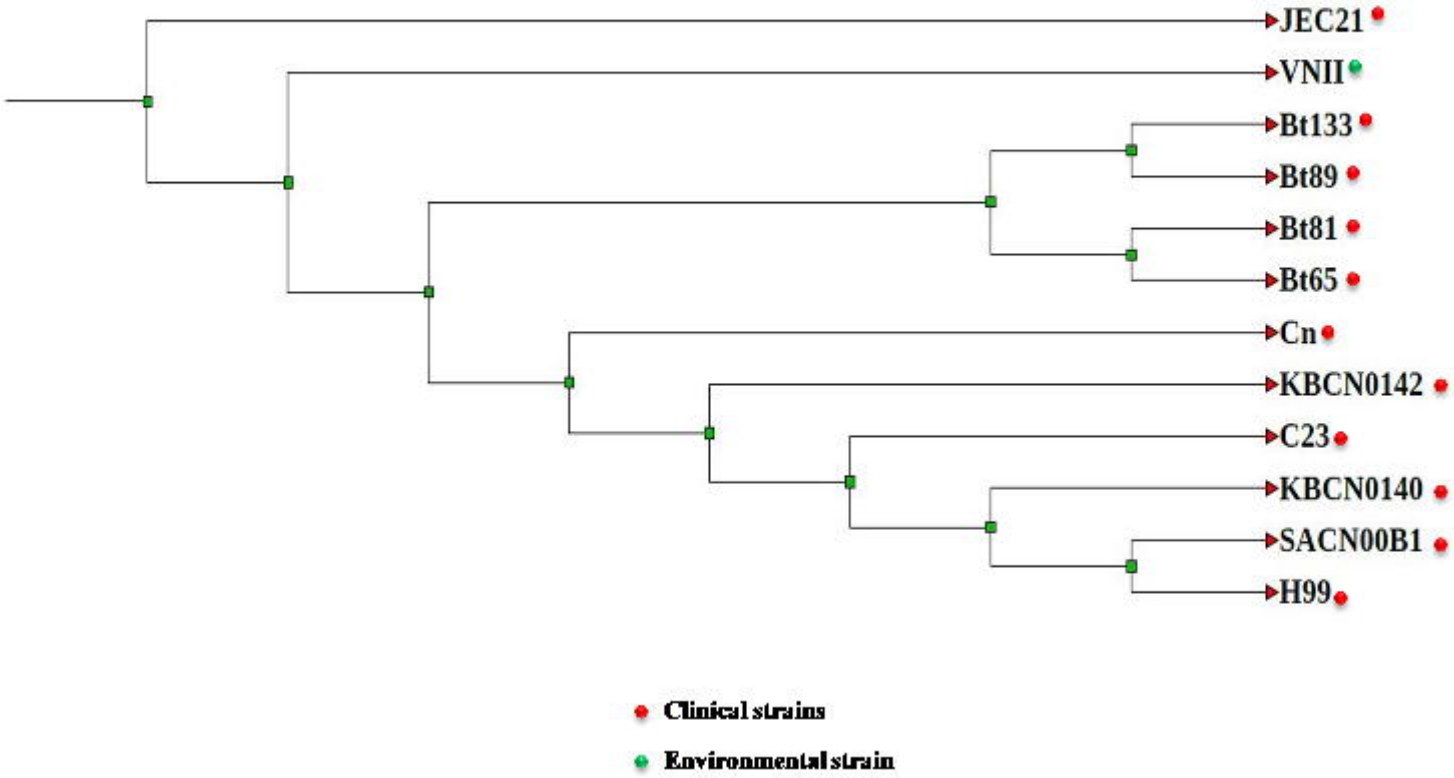
(a)

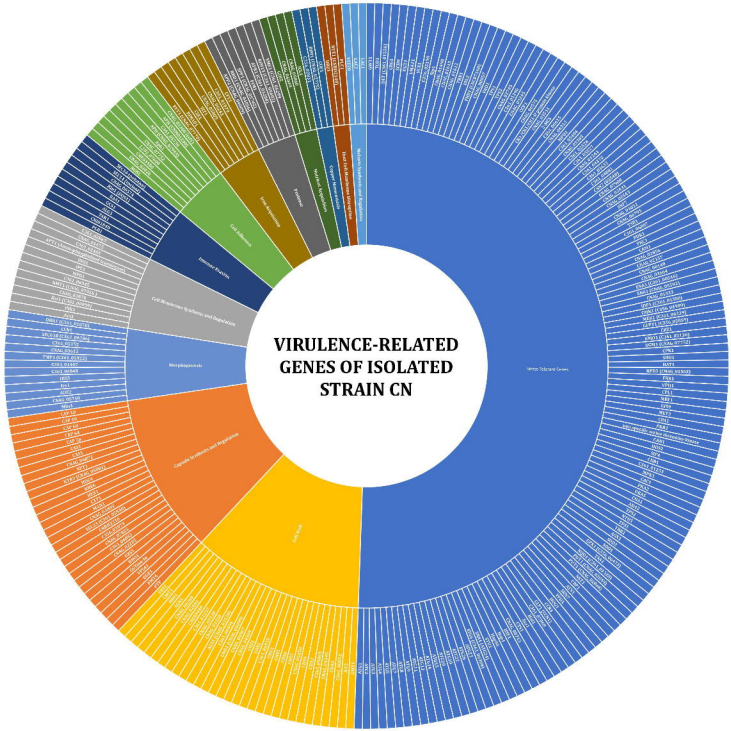


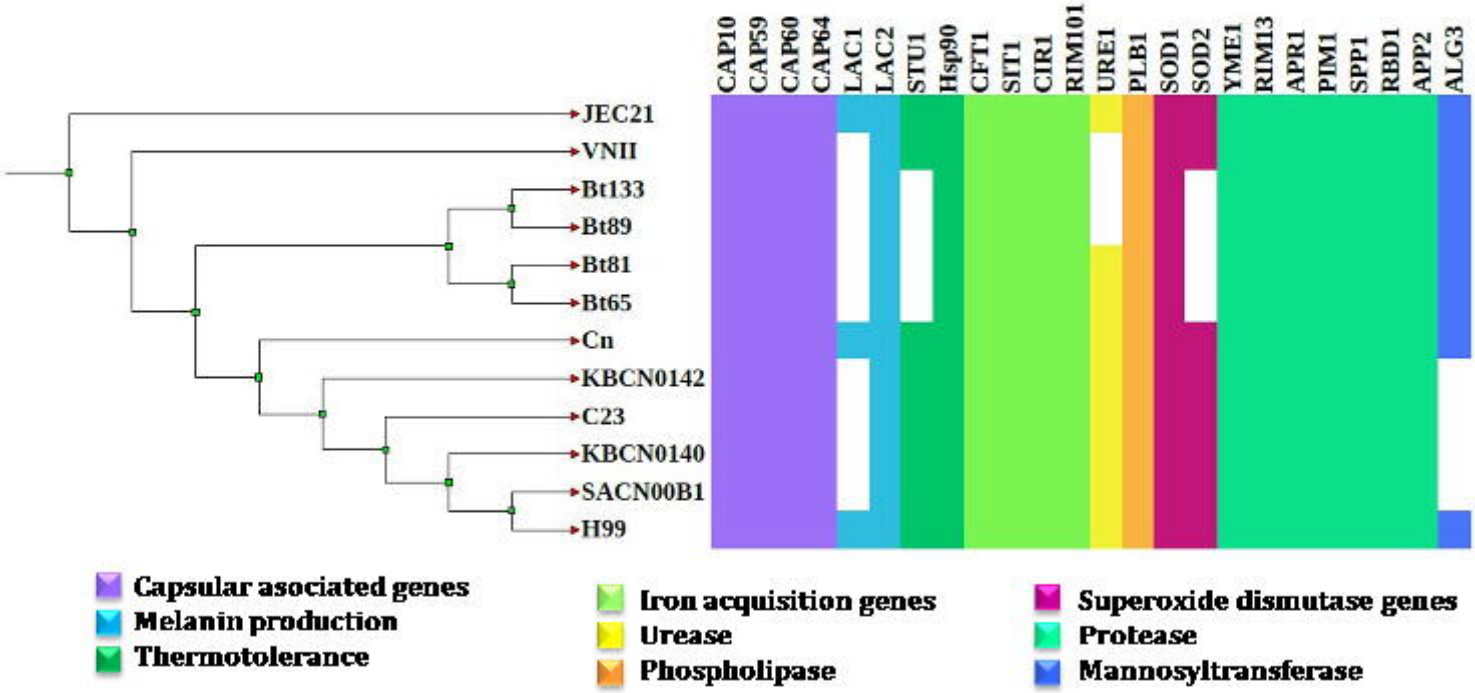
(b)

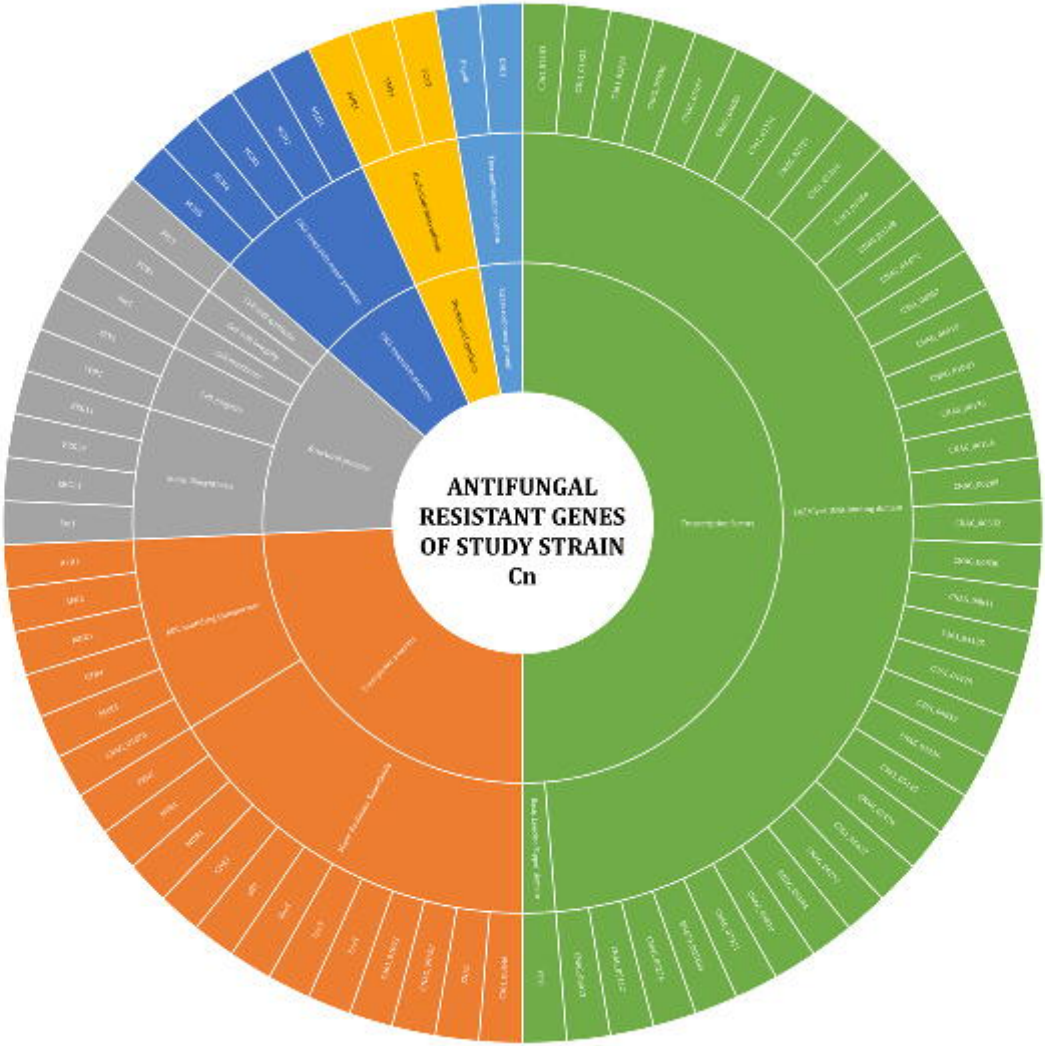

4000000
6000000
8000000
10000000
12000000
14000000
16000000
18000000
SO_10477_Genome_Assembly.fasta
2000000
4000000
6000000
8000000
10000000
12000000
14000000
16000000
18000000

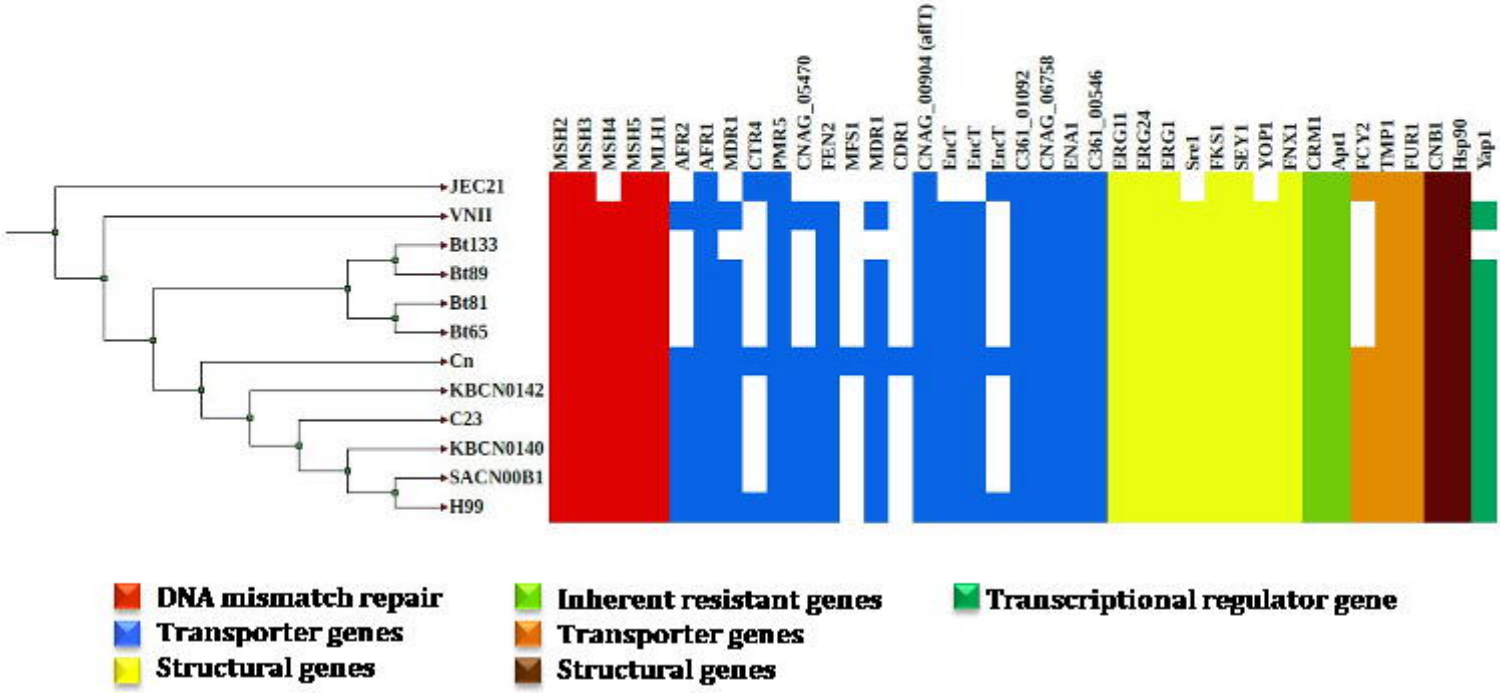
(c)

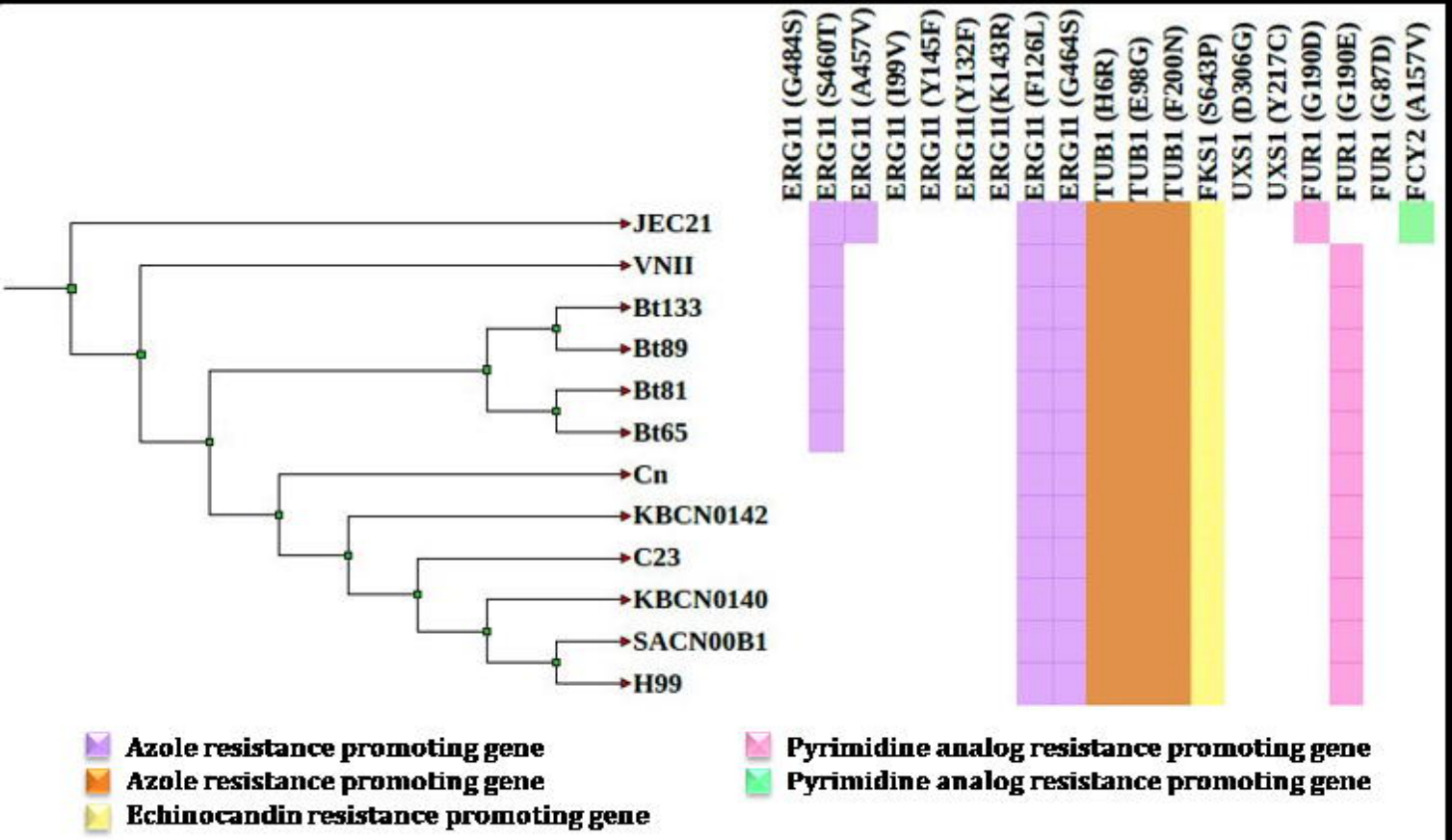

JEC21
VNII
Bt133
Bt89
Bt81
Bt65
Cn
KBCN0142
C23
KBCN0140
SACN00B1
H99
Clinical strains
Environmental strain

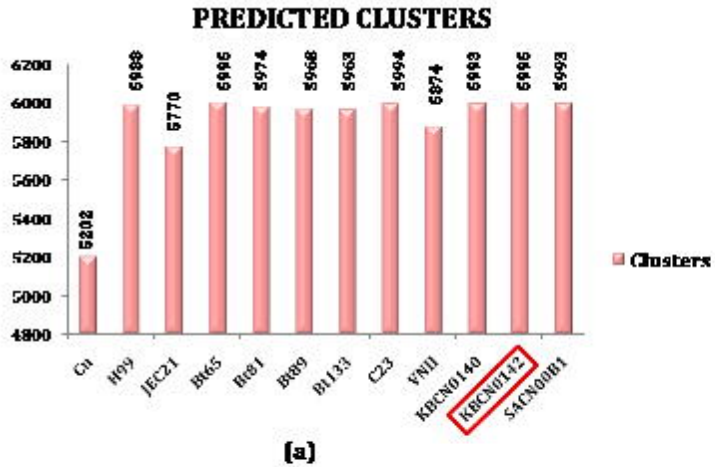

VIRULENCE-RELATED GENES OF ISOLATED STRAIN CN
Vertex Tolerant Genes
Lobes
Capsule Synthesis and Regulation
Morphogenesis
Cell Wall Synthesis and Regulation
Immune Reaction
Cell Wall Synthesis
Transfer
Nutrient Supplement
Cell Wall Synthesis and Regulation
Nutrient Supplement
Cell Wall Synthesis and Regulation
Vertex Tolerant Genes

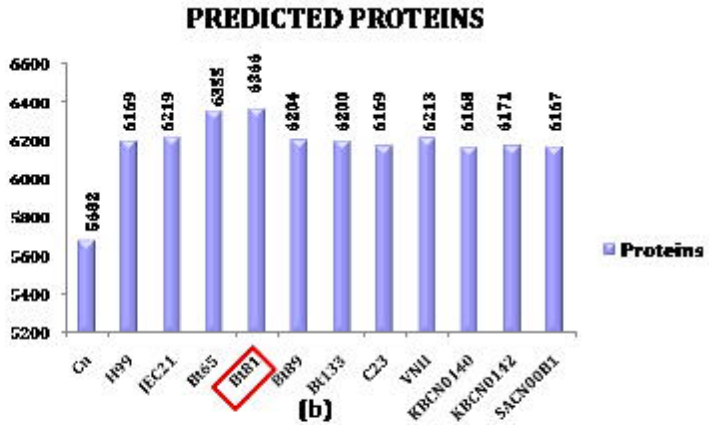

CAP10
CAP59
CAP60
CAP64
LAC1
LAC2
STU1
Hsp90
CFT1
SIT1
CIR1
RIM101
URE1
PLB1
SOD1
SOD2
YME1
RIM13
APR1
PIM1
SPP1
RBD1
APP2
ALG3
JEC21
VNII
Bt133
Bt89
Bt81
Bt65
Cn
KBCN0142
C23
KBCN0140
SACN00B1
H99
Capsular asociated genes
Melanin production
Thermotolerance
Iron acquisition genes
Urease
Phospholipase
Superoxide dismutase genes
Protease
Mannosyltransferase

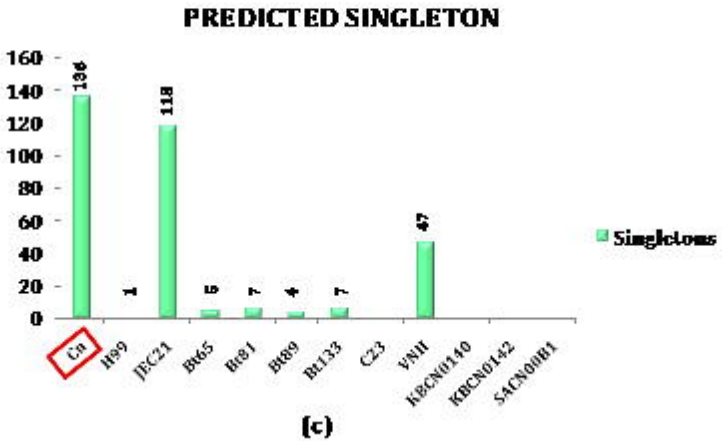

ANTIFUNGAL
RESISTANT GENES
OF STUDY STRAIN
Cn

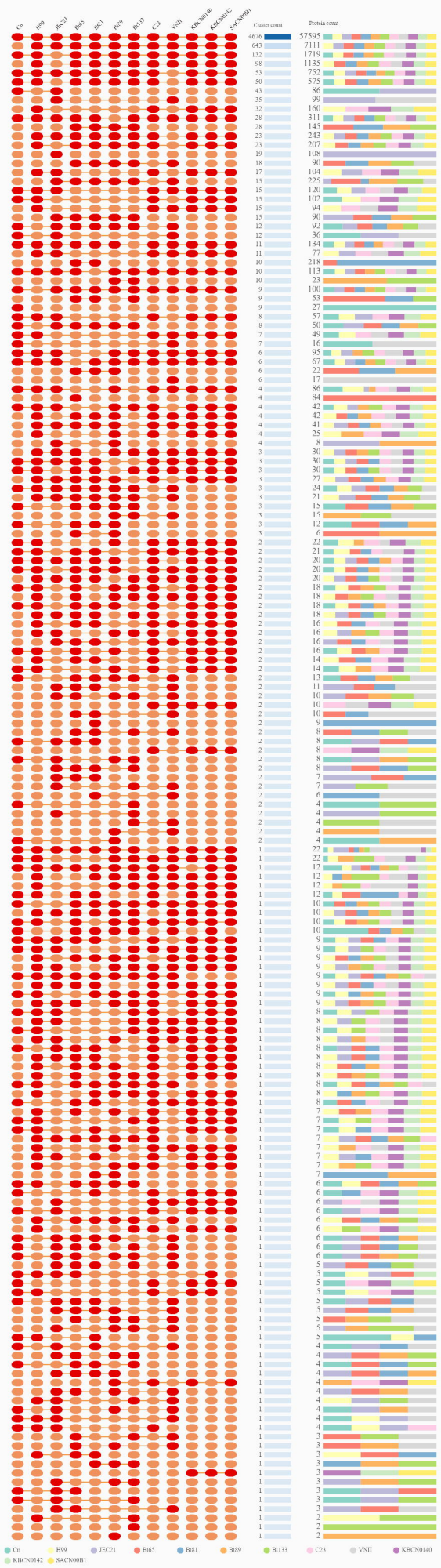

MSH2
MSH3
MSH4
MSH5
MLH1
AFR2
AFR1
MDR1
CTR4
PMR5
CNAG_05470
FEN2
MFS1
MDR1
CDR1
CNAG_00904 (affT)
EncT
EncT
EncT
C361_01092
CNAG_06758
ENA1
C361_00546
ERG11
ERG24
ERG1
Sre1
FKS1
SEY1
YOP1
FNX1
CRM1
Apt1
FCY2
TMP1
FUR1
CNB1
Hsp90
Yap1
JEC21
VNII
Bt133
Bt89
Bt81
Bt65
Cn
KBCN0142
C23
KBCN0140
SACN00B1
H99
DNA mismatch repair
Transporter genes
Structural genes
Inherent resistant genes
Transporter genes
Structural genes
Transcriptional regulator gene

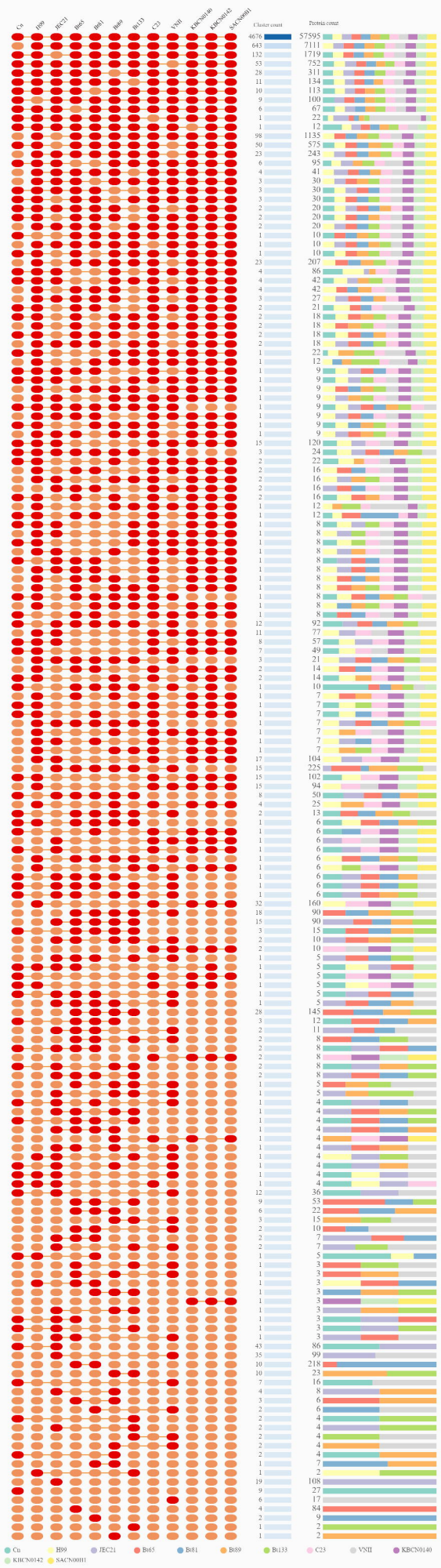

ERG11 (G484S)
ERG11 (S460T)
ERG11 (A457V)
ERG11 (I99V)
ERG11 (Y145F)
ERG11(Y132F)
ERG11(K143R)
ERG11 (F126L)
ERG11 (G464S)
TUB1 (H6R)
TUB1 (E98G)
TUB1 (F200N)
FKS1 (S643P)
UXS1 (D306G)
UXS1 (Y217C)
FUR1 (G190D)
FUR1 (G190E)
FUR1 (G87D)
FCY2 (A157V)
JEC21
VNII
Bt133
Bt89
Bt81
Bt65
Cn
KBCN0142
C23
KBCN0140
SACN00B1
H99
Azole resistance promoting gene
Azole resistance promoting gene
Echinocandin resistance promoting gene
Pyrimidine analog resistance promoting gene
Pyrimidine analog resistance promoting gene

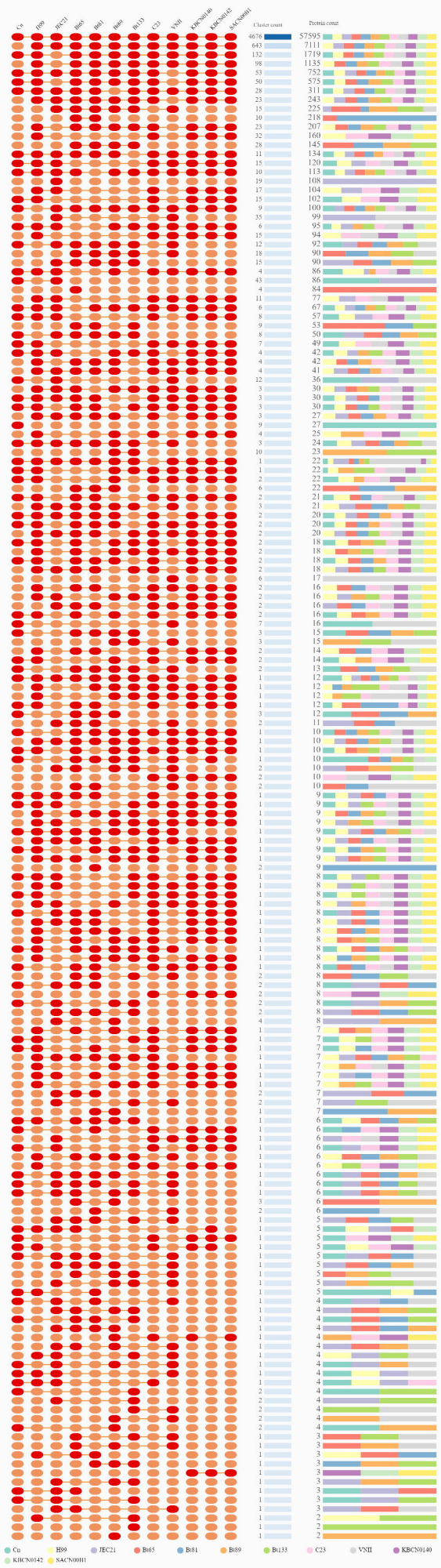

PREDICTED CLUSTERS
6200
6000
5800
5600
5400
5200
5000
4800
5202
5980
5770
5995
5974
5968
5963
5994
5874
5993
5995
5993
Ca
H99
JEC21
B65
Bt81
B89
Bt133
C23
VNII
KBCN0140
KBCN0142
SACN00B1
Clusters
(a)

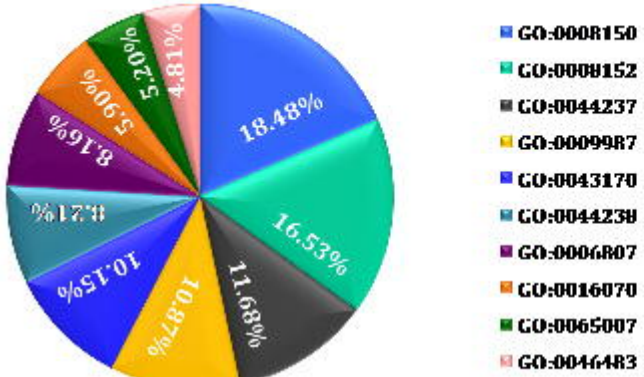

PREDICTED PROTEINS
6600
6400
6200
6000
5800
5600
5400
5200
5602
6169
6219
6355
6366
6204
6200
6169
6213
6168
6171
6167
Ca
H99
JEC21
B65
Bt81
B89
Bt133
C23
VNII
KBCN0140
KBCN0142
SACN00B1
Proteins
(b)


PREDICTED SINGLETON
160
140
120
100
80
60
40
20
0
136
1
118
5
7
4
7
47
Ca
H99
JEC21
B65
Bt81
B89
Bt133
C23
VNII
KBCN0140
KBCN0142
SACN00B1
Singletons
(c)

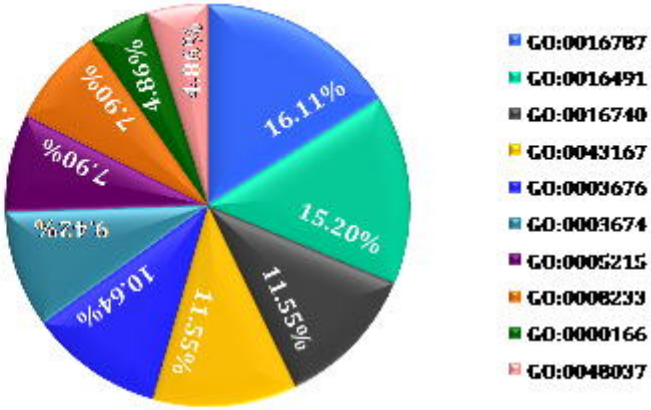

Cu
H99
JEC21
Bi65
Bi81
Bi89
Bi133
C23
VNII
KBCN0140
KBCN0142
SACN00081
Cluster count
Protein count
Cu
H99
JEC21
Bi65
Bi81
Bi89
Bi133
C23
VNII
KBCN0140
KBCN0142
SACN00081


Cu
H99
JEC21
Bi65
Bi81
Bi89
Bi133
C23
VNII
KBCN0140
KBCN0142
SACN00011
Cluster count
Protein count
4676 57595
643 7111
132 1179
53 752
28 311
11 134
10 113
9 100
6 67
1 22
1 12
98 1135
50 575
23 243
6 95
4 41
3 30
3 30
3 30
2 20
2 20
2 20
1 10
1 10
1 10
23 207
4 86
4 42
4 42
3 27
2 21
2 18
2 18
2 18
2 18
1 22
1 12
1 9
1 9
1 9
1 9
1 9
1 9
1 9
15 120
3 24
2 22
2 16
2 16
2 16
1 12
1 12
1 8
1 8
1 8
1 8
1 8
1 8
1 8
1 8
1 8
1 8
12 92
11 77
8 57
7 49
3 21
2 14
2 14
1 10
1 7
1 7
1 7
1 7
1 7
1 7
17 104
15 225
15 102
15 94
8 50
4 25
2 13
1 6
1 6
1 6
1 6
1 6
1 6
1 6
1 6
1 6
32 160
18 90
15 90
3 15
2 10
2 10
1 5
1 5
1 5
1 5
1 5
1 5
28 145
3 12
2 11
2 8
2 8
2 8
2 8
2 8
1 5
1 5
1 4
1 4
1 4
1 4
1 4
1 4
1 4
1 4
1 4
1 4
1 4
12 36
9 53
6 22
3 15
2 10
2 7
2 7
1 5
1 3
1 3
1 3
1 3
1 3
1 3
1 3
1 3
1 3
43 86
35 99
10 218
10 23
7 16
4 8
3 6
2 6
2 4
2 4
2 4
2 4
2 4
1 7
1 2
19 108
9 27
6 17
4 84
2 9
1 2
1 2
Cu
H99
JEC21
Bi65
Bi81
Bi89
Bi133
C23
VNII
KBCN0140
KBCN0142
SACN00011

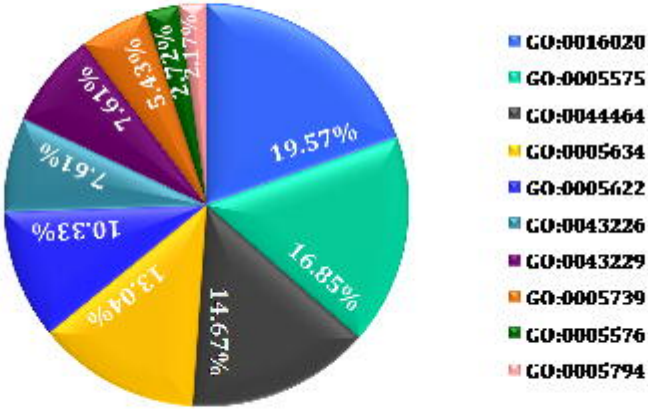

Cu
H99
JEC21
Bi65
Bi81
Bi89
Bi133
C23
VNII
KBCN0140
KBCN0142
SACN00811
Cluster count
Protein count
Cu
H99
JEC21
Bi65
Bi81
Bi89
Bi133
C23
VNII
KBCN0140
KBCN0142
SACN00811


18.48%
16.53%
11.68%
10.87%
10.15%
8.21%
8.16%
5.90%
5.20%
4.81%
GO:0008150
GO:0008152
GO:0044237
GO:0009987
GO:0043170
GO:0044238
GO:0006807
GO:0016070
GO:0065007
GO:0046483

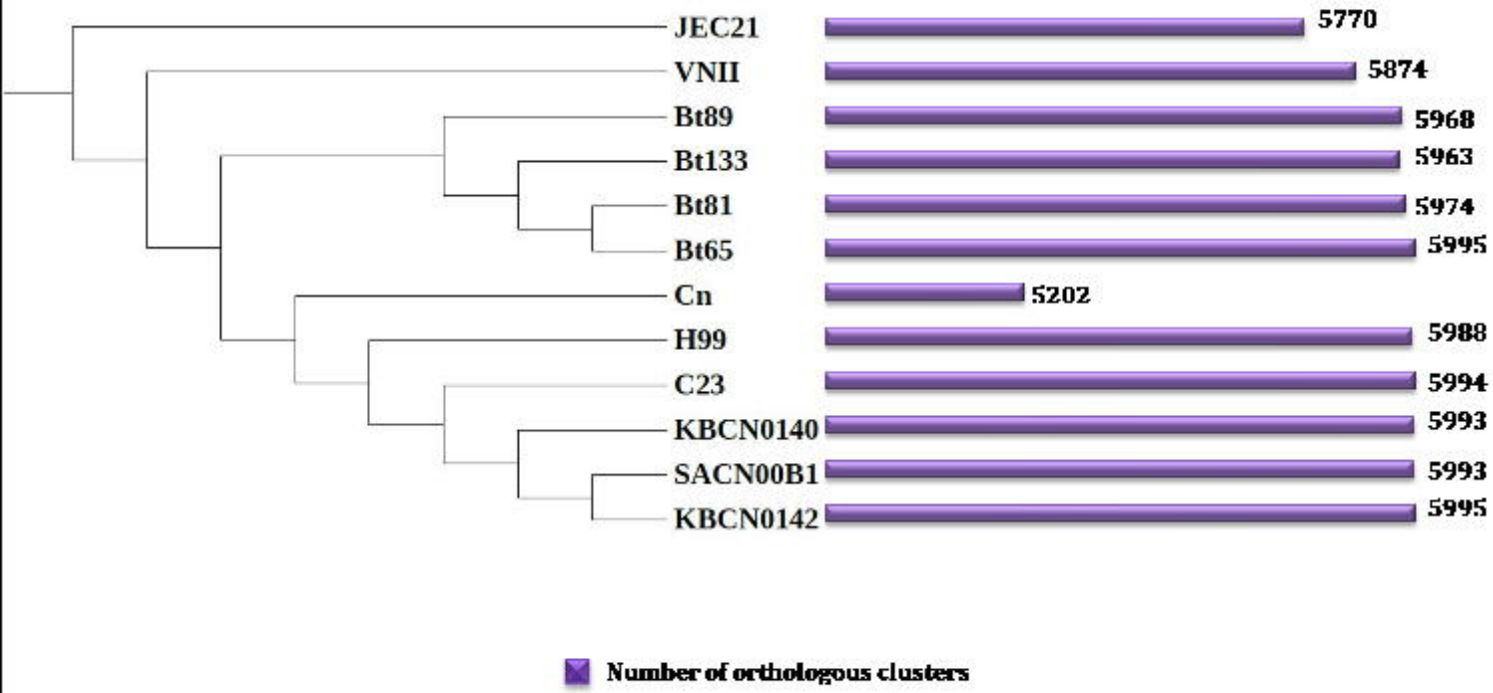
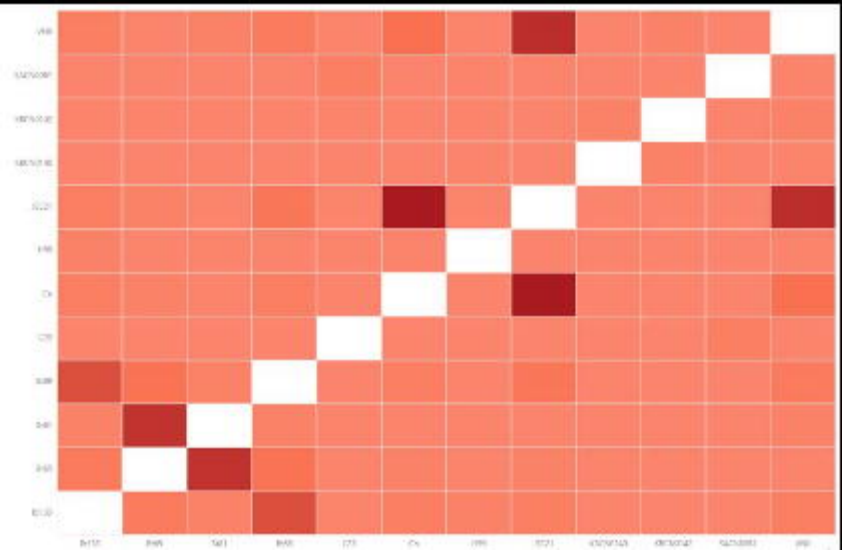
(a)


16.11%
15.20%
11.55%
11.55%
10.64%
9.42%
7.90%
7.90%
4.86%
4.86%
GO:0016787
GO:0016491
GO:0016740
GO:0043167
GO:0003676
GO:0003674
GO:0005215
GO:0008233
GO:0000166
GO:0048037

(b)


19.57%
16.85%
14.67%
13.04%
10.33%
7.61%
7.61%
5.43%
2.72%
2.17%
GO:0016020
GO:0005575
GO:0044464
GO:0005634
GO:0005622
GO:0043226
GO:0043229
GO:0005739
GO:0005576
GO:0005794

(c)


JEC21
VNII
Bt89
Bt133
Bt81
Bt65
Cn
H99
C23
KBCN0140
SACN00B1
KBCN0142
5770
5874
5968
5963
5974
5995
5202
5988
5994
5993
5993
5995
Number of orthologous clusters