The Genome of the Basidiomycetous Yeast and Human Pathogen *Cryptococcus neoformans*

Brendan J. Loftus[1,*], **Eula Fung**[2], **Paola Roncaglia**[3], **Don Rowley**[2], **Paolo Amedeo**[1], **Dan Bruno**[2], **Jessica Vamathevan**[1], **Molly Miranda**[2], **Iain J. Anderson**[1], **James A. Fraser**[4], **Jonathan E. Allen**[1], **Ian E. Bosdet**[5], **Michael R. Brent**[6], **Readman Chiu**[5], **Tamara L. Doering**[7], **Maureen J. Donlin**[8], **Cletus A. D'Souza**[9], **Deborah S. Fox**[4,10], **Viktoriya Grinberg**[1], **Jianmin Fu**[11], **Marilyn Fukushima**[2], **Brian J. Haas**[1], **James C. Huang**[4], **Guilhem Janbon**[12], **Steven J. M. Jones**[5], **Hean L. Koo**[1], **Martin I. Krzywinski**[5], **June K. Kwon-Chung**[13], **Klaus B. Lengeler**[4,14], **Rama Maiti**[1], **Marco A. Marra**[5], **Robert E. Marra**[4,15], **Carrie A. Mathewson**[5], **Thomas G. Mitchell**[4], **Mihaela Pertea**[1], **Florenta R. Riggs**[1], **Steven L. Salzberg**[1], **Jacqueline E. Schein**[5], **Alla Shvartsbeyn**[1], **Heesun Shin**[5], **Martin Shumway**[1], **Charles A. Specht**[16], **Bernard B. Suh**[17], **Aaron Tenney**[6], **Terry R. Utterback**[18], **Brian L. Wickes**[11], **Jennifer R. Wortman**[1], **Natasja H. Wye**[5], **James W. Kronstad**[9], **Jennifer K. Lodge**[8], **Joseph Heitman**[4], **Ronald W. Davis**[2], **Claire M. Fraser**[1], and **Richard W. Hyman**[2]

[1]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

[2]Stanford Genome Technology Center, Stanford University, 855 California Avenue, Palo Alto, CA 94304, USA

[3]Neurobiology Sector, International School for Advanced Studies (SISSA-ISAS), Via Beirut 2-4, 34014 Trieste, Italy

[4]Department of Molecular Genetics and Microbiology, Duke University Medical Center, 322 CARL Building, Research Drive, Box 3546, DUMC, Durham, NC 27710, USA

[5]Genome Sciences Centre, 100-570 West 7th Avenue, Vancouver, BC V5Z 4S6, Canada

[6]Laboratory for Computational Genomics, Washington University, One Brookings Drive, St. Louis, MO 63130, USA

[7]Department of Molecular Microbiology, Washington University School of Medicine, 660 South Euclid Avenue, St. Louis, MO 63110, USA

[8]Department of Biochemistry and Molecular Biology, Saint Louis University School of Medicine, 1402 S. Grand Boulevard, St. Louis, MO 63104, USA

[9]The Michael Smith Laboratories, The University of British Columbia, 2185 East Mall, Vancouver, BC V6T 1Z4, Canada

[10]Research Institute for Children and the Department of Pediatrics, Louisiana State Health Science Center, Children's Hospital, 200 Henry Clay Avenue, New Orleans, LA 70118, USA

[11]University of Texas Health Science Center, 7703 Floyd Curl Drive, San Antonio, TX 78229, USA

---

[*]To whom correspondence should be addressed. bjloftus@tigr.org.

[12]Unité de Mycologie Moléculaire, Institut Pasteur, 25 rue du Docteur Roux, Cedex 15, Paris, France

[13]Molecular Microbiology Section, Laboratory of Clinical Investigation, National Institutes of Health (NIAID/NIH), 9000 Rockville Pike, Bethesda, MD 20892, USA

[14]Institut für Mikrobiologie, Heinrich-Heine-Universität, Universitätsstraße 1/ 26.12, Düsseldorf, Germany

[15]Plant Pathology and Ecology, The Connecticut Agricultural Experiment Station, 123 Huntington Street, New Haven, CT 06511, USA

[16]Department of Medicine, Boston University, 650 Albany Street, EBRC-625, Boston, MA 02118, USA

[17]Department of Biomolecular Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064 USA

[18]Joint Technology Center, J. Craig Venter Foundation, 5 Research Place, Rockville, MD 20850, USA

Abstract

Cryptococcus neoformans is a basidiomycetous yeast ubiquitous in the environment, a model for fungal pathogenesis, and an opportunistic human pathogen of global importance. We have sequenced its ~20-megabase genome, which contains ~6500 intron-rich gene structures and encodes a transcriptome abundant in alternatively spliced and antisense messages. The genome is rich in transposons, many of which cluster at candidate centromeric regions. The presence of these transposons may drive karyotype instability and phenotypic variation. *C. neoformans* encodes unique genes that may contribute to its unusual virulence properties, and comparison of two phenotypically distinct strains reveals variation in gene content in addition to sequence polymorphisms between the genomes.

With an increased immunocompromised population as a result of AIDS and widespread immunosuppressive therapy, *Cryptococcus neoformans* has emerged as a major pathogenic microbe in patients with impaired immunity (1). *C. neoformans* elaborates two specialized virulence factors, a polysaccharide capsule (2) and the antioxidant pigment melanin (3), which enhance human infection and central nervous system colonization. Here, we report the genome sequence of two related strains of *C. neoformans* serotype D (JEC21 and B-3501A) as an important step in the elucidation of the genomic basis for virulence in this pathogenic yeast.

The 19-Mb genome sequence of *C. neoformans* JEC21 [excluding the ribosomal RNA (rDNA) repeats region constituting ~5% of the genome] spans 14 chromosomes from 762 kb to 2.3 Mb (table S1), whereas the 18.5-Mb sequence of the B-3501A strain consists of 14 linked assemblies (scaffolds). Unlike *S. cerevisiae*, the genome of *C. neoformans* shows no evidence for a whole-genome duplication (4). However, a chromosomal translocation and an exact ~60-kb segmental duplication are present in JEC21 compared with B-3501A (5). Almost 5% of the genome consists of transposons, the majority clustered on each chromosome in single blocks that span 40 to 100 kb that may represent sequence-independent regional centromeres, similar to those in *S. pombe* and *N. crassa* (6) (Fig. 1). Each block is unique but all contain at least one copy of the Tcn5 or Tcn6 transposons, which may represent functional elements or target the centromeres. Transposons are also clustered adjacent to the rDNA repeats and within the mating-type (MAT) locus (Fig. 1). In contrast to the other transposons, the long interspersed nuclear element–like (LINE-like) retroelement Cnl1 shows a marked preference for telomeric regions.

To ensure accurate gene structure annotation, sequence data were obtained from both ends of more than 23,000 cDNA clones of a full-length normalized cDNA library from *C. neoformans* JEC21 cells grown under various conditions (7). A total of 6572 protein-encoding genes were identified, which contain an average of 6.3 exons of 255 base pairs (bp) and 5.3 introns of 67 bp (table S2). The mean transcript size of 1.9 kb contains an average of 15% noncoding sequence from both the 5′ and 3′ ends. The gene organization in *C. neoformans* is thus considerably more complex than that of ascomycetes for which genome sequence (table S2) is available and is comparable to that observed in *Arabadopsis thaliana* or *Caenorhabditis elegans*.

A conspicuous feature to emerge from comparing cDNA and genome sequence data is evidence for alternative splicing and endogenous antisense transcripts, in some cases emanating from the same gene locus (Fig. 2). Alternative splicing and natural antisense RNA transcribed in cis were identified in genes encoding diverse functions distributed genome-wide, which suggests that both are widespread genetic regulatory mechanisms in *C. neoformans* (tables S3 to S5). Alternative splice forms were predicted for 277 genes, or 4.2% of the transcriptome (table S4), and a variety of mechanisms could be identified (e.g., exon skipping, truncation, and extension at both 5′ and 3′ ends). Antisense transcripts were identified for 53 genes; however, they appear to have no appreciable coding potential and are usually completely overlapped by their sense counterparts (table S5). The presence and frequency of these antisense transcripts and the presence of the molecular components necessary for RNA interference extend previous studies (8) and indicate that regulation by double-stranded RNA is likely a general regulatory mechanism in this organism.

JEC21 and B-3501A are highly related inbred strains of the alpha mating type, the most prevalent mating type in environmental and clinical isolates (9). As a result of back-crossing during strain construction, the sequence differences that distinguish these strains are restricted to 50% of their genomes, which overall are 99.5% identical at the sequence level. The predicted single-nucleotide polymorphisms (SNPs) and insertion and deletion polymorphisms (indels) are distributed in blocks of high and low sequence polymorphism, reflecting the recombination events that occurred during production of these sibling strains (Fig. 1). The phenotypes of JEC21 and B-3501 differ markedly, with B-3501A being more thermotolerant and more virulent in animal models than JEC21. To investigate the genetic basis for these differences, genomic regions encompassing JEC21 genes were compared directly with the B-3501A assembly. The vast majority (99.7%) of genes share >98% nucleotide identity (fig. S1). Strain-specific genes were experimentally verified by polymerase chain reaction and included a Ras guanosine triphosphatase–activating protein and two proteins of unknown function specific to B-3501A, whereas four proteins of unknown function were specific to JEC21. These genes, in addition to 22 duplicated genes in JEC21 located on the ~60-kb segmental duplication, delineate the strains.

A remarkable feature of *C. neoformans* is the link between virulence and mating type, which is governed by a specialized genomic region, the MAT locus (10). Genome analysis revealed several additional genes in MAT. Numerous other genes involved in mating are not in MAT or on the MAT chromosome and are scattered throughout the genome. Consistent with classification as a heterothallic fungus that does not switch mating type, there are no silent mating-type cassettes.

The major virulence factor of *C. neoformans* is its extensive polysaccharide capsule, an elaborate and dynamic structure that surrounds the fungal cell wall that is unique among fungi that affect humans (2). Genome analysis identified more than 30 new genes likely involved in capsule biosynthesis, including a family containing seven members of the capsule-associated (*CAP64*) gene. The *CAP64* family appears to be restricted to

basidiomycetes, and two members encode alternatively spliced forms (table S5). A second family of six capsule-associated (*CAP10*) genes appears restricted to a subset of fungi and is absent from other yeasts.

The cell wall is an essential and unique component of fungi, and most of the genes involved in the biosynthesis of cell-wall polysaccharides are conserved between the ascomycetes and *C. neoformans*, making them attractive targets for broad-spectrum antifungal drugs. However, *S. cerevisiae* and *C. neoformans* manifest notable differences in their mechanisms of cell-wall protein association. In *S. cerevisiae*, two major classes of proteins are covalently bound to the cell wall: the Pir proteins and a set of proteins that are covalently attached to the cell wall by a glycosylphosphatidylinositol (GPI) anchor. *C. neoformans* lacks both Pir-related genes and several genes that have been implicated in attachment of the GPI anchors to the β-1,6-glucan in the cell wall (11). Genome analysis also predicts more than 50 extracellular mannoproteins that may be associated with the cell wall, most of which are unique to *C. neoformans*.

The phylum Basidiomycota last shared a common ancestor with the ascomycetes ~900 million years ago, and the two phyla have diverged considerably (12). Overall, 65% of *C. neoformans* genes have conserved sequence homologs in a sampling of completed fungal genomes (table S2), and of these 12% are restricted to the basidiomycete genome *Phanerochaete chrysosporium*. Another 10% appear to be unique to *C. neoformans*, based on the absence of identifiable homologs in the current public databases, whereas the remaining 25% match nonfungal sequences (7). Lineage-specific gene family expansions do not represent the most abundant protein domains within the *C. neoformans* genome, which are similar to those of ascomycetous fungi (tables S6 and S7). Two of the 11 gene families that appear unique to *C. neoformans* are involved in capsule formation, and another encodes nucleotide sugar epimerases associated with cell-wall formation. About 60% of the *C. neoformans* genes could be assigned gene ontology terms for molecular function (7), and comparison with *S. cerevisiae* reveals a similar distribution of genes across nearly all functional categories (fig. S2). One exception is an expansion of the drug-efflux transporters of the major facilitator superfamily in *C. neoformans*, which suggests enhanced transport capability in this environmental yeast.

Recently, the *Candida albicans* genome was reported (13), enabling a comparison between these divergent pathogenic fungi. *C. neoformans* is an environmental organism that infects through inhalation, whereas *C. albicans* is part of normal human microbiota and infects by bloodstream invasion. Myriad cell-surface proteins implicated in *C. albicans* adhesion to epithelial cells are absent in *C. neoformans*, which suggests that *C. neoformans* binds host cells by distinct mechanisms. *C. neoformans* elaborates both capsule and melanin; *C. albicans* makes neither and lacks genes for their production.

The *C. neoformans* genome sequence provides new insights into this important fungal human pathogen. The genome encodes a core complement of genes common to other fungi and, despite a large divergence time, the functional distribution of many *C. neoformans* genes mirrors that of *S. cerevisiae*. By contrast with *S. cerevisiae*, however, the *C. neoformans* genome displays an intron-rich gene tapestry and a transcriptome rife with alternative splicing and antisense transcripts. These genome sequence data, together with those from another basidiomycete, *P. chrysosporium* (14), suggest that more complex gene structures may be a general feature of basidiomycetes (table S2). The genome sequence data described herein from two closely related strains of *C. neoformans* provide a foundation to explore the molecular basis of virulence in this pathogen and reveal differences in virulence strategies between *C. neoformans* and other pathogenic fungi.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References and Notes

1. Casadevall, A.; Perfect, JR. Cryptococcus neoformans. ASM Press; Washington, DC: 1998.

2. Bose I, Reese AJ, Ory JJ, Janbon G, Doering TL. Eukaryot Cell. 2003; 2:655. [PubMed: 12912884]

3. Casadevall A, Rosas AL, Nosanchuk JD. Curr Opin Microbiol. 2000; 3:354. [PubMed: 10972493]

4. Kellis M, Birren BW, Lander ES. Nature. 2004; 428:617. [PubMed: 15004568]

5. Fraser JA, et al. in preparation.

6. Cambareri EB, Aisner R, Carbon J. Mol Cell Biol. 1998; 18:5465. [PubMed: 9710630]

7. Materials and methods are available as supporting material on *Science* Online.

8. Gorlach JM, McDade HC, Perfect JR, Cox GM. Microbiol. 2002; 148:213.

9. Kwon-Chung KJ, Bennett JE. Am J Epidemiol. 1978; 108:337. [PubMed: 364979]

10. Lengeler KB, et al. Eukaryot Cell. 2002; 1:704. [PubMed: 12455690]

11. Shahinian S, Bussey H. Mol Microbiol. 2000; 35:477. [PubMed: 10672173]

12. Hedges SB, Blair JE, Venturi ML, Shoe JL. BMC Evol Biol. 2004; 4:2. [PubMed: 15005799]

13. Jones T, et al. Proc. Natl. Acad. Sci. U.S.A. 2004; 101:7329.

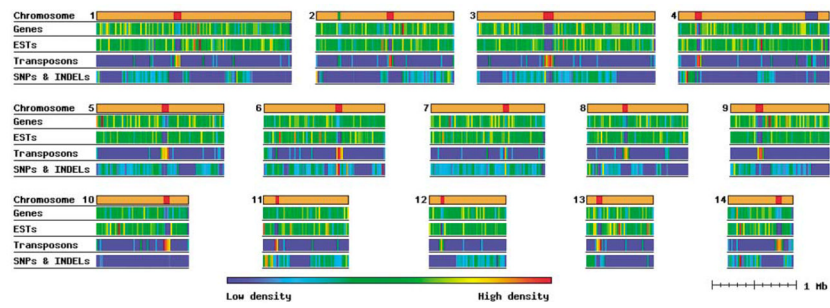14. Martinez D, et al. Nature Biotechnol. 2004; 22:695. [PubMed: 15122302]

Fig. 1.
The *C. neoformans* JEC21 genome with each chromosome represented as a colored bar. Specific features are pseudocolored, from red (high density) to deep blue (low density) and plotted on a log scale. These include the density of genes, transposons, expressed sequence tags (ESTs), and predicted SNPs and indels. Candidate centromeric regions and the MAT locus are represented as red bars and a blue bar, respectively. The location of the rDNA repeat is represented by a green bar.
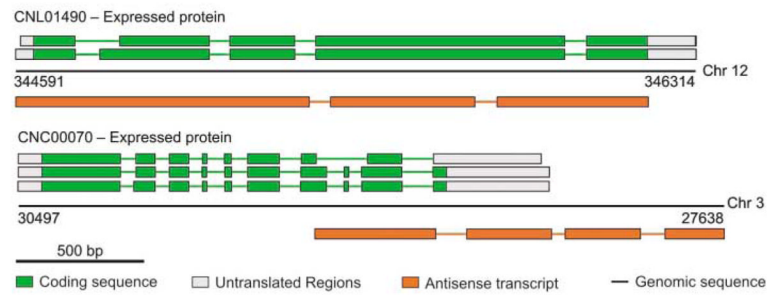
Fig. 2.
Gene structures that display evidence for both alternative splicing and natural in cis antisense transcripts based on JEC21 cDNA alignments to the genome sequence. Colored boxes represent exonic regions. Each gene structure represents an alternative spliced form. The black line represents the genomic sequence.