Comparative gene genealogical analyses of strains of serotype AD identify recombination in populations of serotypes A and D in the human pathogenic yeast Cryptococcus neoformans

Jianping Xu¹ and Thomas G. Mitchell²

¹Department of Biology, McMaster University, Hamilton, Ontario, Canada L8S 4K1
²Department of Molecular Genetics and Microbiology, Duke University Medical Center, Durham, NC 27710, USA

Cryptococcus neoformans is a major pathogen of humans throughout the world. Using commercial monoclonal antibodies to capsular epitopes, strains of C. neoformans manifest five serotypes: A, B, C, D and AD. Previous studies demonstrated significant divergence among serotypes A, B, C and D, which are typically haploid. In contrast, most strains of serotype AD are diploid or aneuploid and result from recent hybridization between strains of serotypes A and D. Whether serotypes A, B, C and D represent strictly asexual lineages is not known. Using comparative genealogical analyses of two genes, the authors investigated whether recombination occurred among strains within serotypes A and D. For each of 14 serotype AD strains, a portion (642 bp) of the orotidine monophosphate pyrophosphorylase (URA5) gene was cloned and sequenced. Each of these 14 strains contained two different alleles and sequences for both alleles were obtained. The URA5 gene genealogy was compared to that derived from the laccase (LAC) gene, which was reported recently for the same 14 strains. For both genes, each of the 14 serotype AD strains contained two phylogenetically distinct alleles: one allele was highly similar to those from serotype A strains and the other to alleles from serotype D strains. However, within both the serotype A allelic group and the serotype D allelic group, there was significant incongruence between genealogies derived from URA5 and LAC. The results suggest recombination in natural populations of both serotypes A and D.

INTRODUCTION

Cryptococcus neoformans is an important fungal pathogen of humans and other mammals throughout the world (Mitchell & Perfect, 1995). Strains of C. neoformans are traditionally divided into five distinct serotypes: A, B, C, D and AD (see summary in Casadevall & Perfect, 1998). Recent studies show that strains of different serotypes often exhibit significant divergence at the molecular level (e.g. Xu et al., 2000b). Three varieties have been proposed to accommodate this divergence: C. neoformans var. neoformans, representing serotype D strains; C. neoformans var. grubii to represent serotype A strains; and C. neoformans var. gattii for serotype B and C strains (Kwon-Chung et al., 1982; Franzot et al., 1999). Recently, based on evidence from amplified fragment length polymorphisms (AFLP), Boekhout et al. (2001) proposed to divide C. neoformans into two separate species, with serotypes A, D and AD included in the species Cryptococcus neoformans (Sanfelice) Vuillemin, and strains of serotypes B and C constituting a new species, Cryptococcus bacillisporus Kwon-Chung. Unlike the previous classification with three varieties (Franzot et al., 1999), varieties were not included in this two-species proposal (Boekhout et al., 2001). Because the nomenclature is still unsettled but nevertheless based on serotype identification, to avoid confusion, only the serotype will be used to describe the strains in this study.

C. neoformans is a basidiomyceteous yeast, typically with a haploid nucleus. It has a heterothallic life cycle. The mating type locus has two alternative alleles, a and a. Under suitable conditions, yeast cells with opposite mating types can fuse to form dikaryotic hyphae. The terminal cell of a dikaryotic hypha forms a basidium where nuclear fusion and meiosis occur, and where four chains of haploid basidiospores are produced (Kwon-Chung, 1976). Most current genetic analyses of C. neoformans involve strains of serotype D, as mating between strains of opposite mating types within this serotype occurs readily and karyogamy, meiosis and basidiospore formation proceed normally. However, when strains of serotypes A and D mate in the laboratory, meiotic

Abbreviations: ILD, incongruence-length difference; MP, maximum parsimony; PH, partition homogeneity.

Received 11 December 2002
Revised 30 April 2003
Accepted 6 May 2003
progeny often contain alleles from both parental strains and are diploid or aneuploid (e.g. Lengeler et al., 2000). Recent molecular analyses have shown that most environmental and clinical strains of serotype AD are diploid or aneuploid and contain alleles typical of both serotypes A and D (Brandt et al., 1993, 1995; Boekhout et al., 2001; Lengeler et al., 2000; Xu et al., 2002). In addition, genealogical analyses indicated that strains of serotype AD are recent hybrids between strains of serotypes A and D and involved multiple hybridization events (Xu et al., 2002). As in a previous study (Xu et al., 2002), the terms 'hybridization' and 'hybrid' denote the process or a product of the process, respectively, by which an offspring is generated from a mating between two genetically divergent parental strains (e.g. strains of different breeds, varieties, subspecies, species or genera). Because strains of serotypes A and D diverged millions of years ago (Xu et al., 2000b), mating between strains of serotypes A and D is referred to as hybridization, comparable to the terminology applied to plants and animals. In contrast, recombination is generally defined as the formation of new combinations of genes in progeny that did not exist in the parents, resulting from the processes of mating, crossing-over, and independent assortment among genes located on different chromosomes during meiosis.

Despite extensive population surveillance and strain typing studies of C. neoformans, several questions regarding its population biology remain unresolved. One question is whether recombination occurs in populations within a serotype. Current evidence indicates that there are clones and clonal lineages within each of the serotypes, consistent with extensive asexual reproduction, clonal dispersal and lack of recombination. For example, epidemiological surveys in four areas of the US revealed only a few multilocus enzyme electrophoretic genotypes, and most genotypes were represented by multiple strains that were often distributed over wide geographical areas (Brandt et al., 1996).

The extensive clonality observed in natural populations of C. neoformans might reflect a bias in sampling. More than 90% of all clinical and environmental isolates of C. neoformans of serotype D possess the α mating type (MATα), and nearly 100% of isolates of serotype A are MATα. For example, among 355 clinical and environmental strains of serotype A, D or AD, which were obtained from four areas in the USA, only one strain contained only the mating type α (MATα) allele at the STE20 gene, which is located within the mating type region, and this strain had serotype AD (Yan et al., 2002). Clinical isolates of serotype A or D strains are predominantly MATα, probably because pathogenicity is linked to this locus (Kwon-Chung et al., 1992). The predominance of MATα among environmental isolates may similarly reflect a relationship of this locus to fitness, although there is no experimental evidence to support this hypothesis. The second hypothesis is related to haploid fruiting ability of MATα strains. Under starvation conditions in the laboratory, MATα strains produce hyphae, basidia and haploid fruiting (Wickes et al., 1996). Because of their small size, basidiospores are more effectively dispersed than the encapsulated vegetative yeast cells. Thus, prolific haploid fruiting of MATα strains may explain the dominance of this mating type among environmental isolates that might also engage in true sexual reproduction in the presence of both mating types.

In contrast to the biased mating type ratios in clinical and environmental isolates of serotypes A and D, mating type ratios in strains of serotype AD are generally balanced. This is because most strains of serotype AD possess both MATα and MATα loci (Lengeler et al., 2001; Yan et al., 2002). In addition, in one study, a significant proportion (14 of 19) of serotype AD isolates contained the MATα allele from serotype A parents (Aa) (Yan et al., 2002). Strains of Aa were previously thought to be extinct and were identified only recently (Lengeler et al., 2000; Viviani et al., 2001). Because of the balanced ratios of mating type alleles among AD strains and the existence of strains with mating type allele Aa, we decided to test strains of serotype AD for the possibility of recombination for serotypes A and D.

There are several methods to detect recombination in natural populations of microorganisms (Xu & Mitchell, 2003). One method is to compare genealogies of different genes for the same set of strains. In strictly asexual organisms with no sexual recombination, genealogies of different genes are expected to be congruent. Conversely, incongruent gene genealogies would be consistent with recombination. In a previous study (Xu et al., 2002), we analysed a portion of the LAC gene sequence from each of 14 serotype AD strains. These 14 strains were all from the USA, and 12 were collected from the San Francisco area. In this study, we sequenced a portion (642 nucleotides) of another gene, orotidine monophosphate pyrophosphorylase (URA5) for the same 14 strains and compared genealogies derived from these two genes. URA5 was selected for study because it is present in only one copy in haploid strains of C. neoformans and known to be highly polymorphic (Edman & Kwon-Chung, 1990; Casadevall et al., 1992; Xu et al., 2000b). To strengthen the analyses, we included additional strains of serotypes A and D. The results revealed gene genealogy incongruence, consistent with sexual recombination within populations of both serotypes A and D.

METHODS

Strains. The 14 isolates analysed in this study were the same as those described by Xu et al. (2002) and were collected from three geographical areas in the USA during a population-based active surveillance conducted by the Centers for Disease Control and Prevention in Atlanta, GA, USA (Brandt et al., 1995, 1996). These 14 strains included all three multilocus enzyme electrophoretic (MLEE) genotypes detected for serotype AD strains (Brandt et al., 1993, 1995). Using PCR primers that were specific for serotype-specific mating types (Yan et al., 2002), we determined the serotype and mating type alleles of these 14 clinical isolates. Table 1 presents their geographical origins, dates of isolation, and mating type alleles.
Table 1. Serotype AD strains of *C. neoformans* investigated in this study

<table>
<thead>
<tr>
<th>Strain*</th>
<th>Geographical origin, year isolated</th>
<th>Mating type†</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAS92-0022</td>
<td>San Francisco, 1992</td>
<td>AaDz</td>
</tr>
<tr>
<td>MAS92-0086</td>
<td>San Francisco, 1992</td>
<td>AaDz</td>
</tr>
<tr>
<td>MAS92-0153</td>
<td>San Francisco, 1992</td>
<td>AaDz</td>
</tr>
<tr>
<td>MAS92-0189</td>
<td>Georgia, 1992</td>
<td>AaDz</td>
</tr>
<tr>
<td>MAS92-0224</td>
<td>San Francisco, 1992</td>
<td>A–Dz</td>
</tr>
<tr>
<td>MAS92-0668</td>
<td>San Francisco, 1992</td>
<td>AaDz</td>
</tr>
<tr>
<td>MAS92-0793</td>
<td>San Francisco, 1992</td>
<td>AzDz</td>
</tr>
<tr>
<td>MAS92-0855</td>
<td>San Francisco, 1992</td>
<td>AzD–</td>
</tr>
<tr>
<td>MAS93-0315</td>
<td>San Francisco, 1993</td>
<td>AaDz</td>
</tr>
<tr>
<td>MAS93-0610</td>
<td>San Francisco, 1993</td>
<td>AaDz</td>
</tr>
<tr>
<td>MAS94-0018</td>
<td>San Francisco, 1994</td>
<td>AaDz</td>
</tr>
<tr>
<td>MAS94-0241</td>
<td>San Francisco, 1994</td>
<td>AaDz</td>
</tr>
<tr>
<td>MAS94-0244</td>
<td>San Francisco, 1994</td>
<td>AaDz</td>
</tr>
<tr>
<td>MAS94-0351</td>
<td>Texas, 1994</td>
<td>A–Da</td>
</tr>
</tbody>
</table>

*For additional information about these strains, see Brandt *et al.* (1993, 1995) and Xu *et al.* (2002).
†Mating type was determined as described by Yan *et al.* (2002).

Strains MAS93-0315 and MAS93-0610 were isolated at different times from cerebrospinal fluid of the same patient. From another patient, strains MAS94-0241 and MAS94-0244 were isolated from a bronchial wash and the prostate gland, respectively. Each of the other strains was isolated from a different patient.

DNA manipulations. DNA was isolated from each strain as described previously (Xu *et al.*, 2000a). The orotidine monophosphate pyrophosphorylase (*URA5*) gene was amplified, cloned, sequenced, and analysed from each strain. The fragment was amplified using the following oligonucleotide primers (5′–3′): forward: ACCGCCTGCTGTATCTA, reverse: GGACATGATGATTGGAGT. The amplified DNA fragment corresponded to the complete nucleotide sequence (1–779 bp) as reported for a serotype D strain by Edman & Kwon-Chung (1990; GenBank accession number M93026). Because ends of sequences were ambiguous on typical sequencing gels, only the unambiguous nucleotides from position 65 to 706 (a total of 642 nucleotides) were analysed for all strains in this study. The analysed fragment contains two introns, one complete exon, and partial sequences from two other exons (see Edman & Kwon-Chung, 1990).

A typical PCR reaction contained 10 µl (~1 ng) of diluted genomic DNA template, 0·5 unit AmpliTag DNA polymerase, 0·2 mM of each primer and 0·2 mM of each deoxyribonucleotide triphosphate in a total volume of 50 µl. The following PCR conditions were used: 3 min at 94 °C, followed by 40 cycles of 30 s at 94 °C, 30 s at 50 °C and 30 s at 72 °C, and lastly, 7 min of extension at 72 °C. PCR products were cleaned by using Wizard spin columns (Promega) and sequenced using an Applied Biosystems Prism 373 (ABI373) automated sequencer with dRhodamine-labelled terminators (PE Applied Biosystems) following the manufacturer’s instructions. However, we were unable to obtain clear *URA5* sequences from any of the 14 strains by the direct PCR sequencing method. Similar to those for the *LAC* gene, this result suggested sequence heterogeneity at the *URA5* locus within each strain. We therefore cloned the PCR product from each strain using a pGEM-T cloning kit (Promega) and transformed the cloned PCR products into competent *Escherichia coli* cells following the manufacturer’s instructions. For each of the 14 strains, ten random *E. coli* colonies were picked, amplified with the *URA5* primers, and digested with the restriction enzyme *Dde*I to screen for different alleles. Here, the term ‘allele’ refers to a distinct nucleotide sequence from a locus within a strain. The term ‘haplotype’ indicates a unique sequence in the whole collection of strains. For each strain, two clones representing each of two *Dde*I restriction digest patterns were sequenced (i.e. four clones total for each strain) to control for DNA sequence variation generated due to PCR, cloning and sequencing. Among the 14 strains, only one (MAS94-0241) showed a single base pair difference between two presumably identical alleles (i.e. two clones with the same *Dde*I restriction digest pattern). To check the authenticity of these two clones, two additional clones were sequenced. The additional sequences were the same as one of the two original clones and identical to the allele from MAS94-0244, a strain from a different body site of the same patient. These sequences were aligned and optimized and imported to the software PAUP* (Swofford, 2002).

Data analyses. Phylogenetic analysis was performed with PAUP* (Swofford, 2002). Maximum parsimony (MP) trees were identified using heuristic searches based on 500 random sequence additions (Swofford, 2002). Statistical support for phylogenetic groupings was assessed by bootstrap analysis using 1000 replicate datasets (sampled from phylogenetically informative characters) with the random addition of sequences during each heuristic search. This analysis identified phylogenetically distinct and statistically well-supported sequence clusters.

The partition homogeneity (PH) test was used to determine statistical significance of congruence between gene genealogies and to infer recombination within populations of serotypes A and D. Because this test requires that the rates of molecular evolution did not differ significantly among lineages, we first determined if the two genes evolved according to a molecular clock model using sequences from the 14 strains of serotype AD. Maximum-likelihood estimates of the MP trees (Felsenstein, 1981) with and without a molecular clock showed little difference (*P* > 0·2 for both genes), demonstrating that the evolution of *LAC* and *URA5* was not significantly different from molecular clock models. The PH test was then used to compare the gene genealogies of the serotype A and serotype D haplotype clusters (see below).

To determine the relationships among haplotypes of strains of serotype AD and those of strains of serotypes A or D, we included published *URA5* sequences from strains of serotypes A and D (Xu *et al.*, 2000b; GenBank accession numbers AF140185–AF140217). For strains of serotypes A, B, C and D (and some strains of serotype AD) sequenced previously, direct sequencing using PCR products without cloning revealed no sequence ambiguity, and each strain had only one allele at each of four genes analysed, including *URA5* (Xu *et al.*, 2000b; see also Edman & Kwon-Chung, 1990; Casadevall *et al.*, 1992). Since introns in both *URA5* and *LAC* genes from strains of serotypes B and C were difficult to align with those from strains of serotypes A and D (Xu *et al.*, 2000b), strains of serotypes B and C were not included in the analyses here.
RESULTS

Sequence variation at the URA5 gene within and among strains
Of the 642 unambiguously aligned nucleotides, 51 were variable among the 28 DNA fragments from the 14 serotype AD strains. Only the 51 polymorphic nucleotide sites are shown in Table 2. The remaining 591 sites were identical among all 28 sequences and therefore offered no discriminating power for haplotype identification. Among the 51 polymorphic sites, 17 were located in introns (see underlined nucleotide positions in Table 2) and 34 in exons. The two introns included a total of 104 aligned nucleotide positions and the exons had 538 nucleotides. Introns for the URA5 gene as described by Edman & Kwon-Chung (1990) correspond to the following nucleotide positions: intron 1, 100–154; and intron 2, 383–431. Not surprisingly, the introns had a higher percentage of polymorphic sites than exons ($\chi^2 = 9.40$, df = 1, $P < 0.01$). Three sites with insertions/deletions were all located within introns.

Among these 28 sequences, 14 unique haplotypes were found (Table 2). These haplotypes were present in different frequencies, with haplotype 2 in four strains, haplotype 8 in two strains, haplotype 11 in five strains, and haplotype 12 in seven strains. Haplotypes 1, 3, 4, 5, 6, 7, 9, 10, 13 and 14 were found in one strain each (Table 2). A total of 10 diploid genotypes were found among the 14 strains at the

<table>
<thead>
<tr>
<th>Strain</th>
<th>Allele</th>
<th>Polymorphic nucleotide position*</th>
<th>Haplotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAS92-0022</td>
<td>1</td>
<td>CATCGCCCAATTCGCTATCTCACTACGAGGACTCGGCTCTCTGACTGCTCTTTA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>GGC-ATGAGGTGTCTGCTGTTCTGTTCTGAAGG</td>
<td>11</td>
</tr>
<tr>
<td>MAS92-0086</td>
<td>1</td>
<td>CATCGCCCAATTCGCTATCTCACTACGAGGACTCGGCTCTCTGACTGCTCTTTA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>GGC-ATGAGGTGTCTGCTGTTCTGTTCTGAAGG</td>
<td>11</td>
</tr>
<tr>
<td>MAS92-0153</td>
<td>1</td>
<td>TGC-ATGAGGTGTCTGCTGTTCTGTTCTGAAGG</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>GGC-ATGAGGTGTCTGCTGTTCTGTTCTGAAGG</td>
<td>11</td>
</tr>
<tr>
<td>MAS92-0189</td>
<td>1</td>
<td>GGC-ATGAGGTGTCTGCTGTTCTGTTCTGAAGG</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>GGC-ATGAGGTGTCTGCTGTTCTGTTCTGAAGG</td>
<td>11</td>
</tr>
<tr>
<td>MAS92-0224</td>
<td>1</td>
<td>CATCGCCCAATTCGCTATCTCACTACGAGGACTCGGCTCTCTGACTGCTCTTTA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>GGC-ATGAGGTGTCTGCTGTTCTGTTCTGAAGG</td>
<td>11</td>
</tr>
<tr>
<td>MAS92-0668</td>
<td>1</td>
<td>CATCGCCCAATTCGCTATCTCACTACGAGGACTCGGCTCTCTGACTGCTCTTTA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>GGC-ATGAGGTGTCTGCTGTTCTGTTCTGAAGG</td>
<td>11</td>
</tr>
<tr>
<td>MAS92-0793</td>
<td>1</td>
<td>CATCGCCCAATTCGCTATCTCACTACGAGGACTCGGCTCTCTGACTGCTCTTTA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>GGC-ATGAGGTGTCTGCTGTTCTGTTCTGAAGG</td>
<td>11</td>
</tr>
<tr>
<td>MAS92-0855</td>
<td>1</td>
<td>CATCGCCCAATTCGCTATCTCACTACGAGGACTCGGCTCTCTGACTGCTCTTTA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>GGC-ATGAGGTGTCTGCTGTTCTGTTCTGAAGG</td>
<td>11</td>
</tr>
<tr>
<td>MAS93-0315</td>
<td>1</td>
<td>CATCGCCCAATTCGCTATCTCACTACGAGGACTCGGCTCTCTGACTGCTCTTTA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>GGC-ATGAGGTGTCTGCTGTTCTGTTCTGAAGG</td>
<td>11</td>
</tr>
<tr>
<td>MAS93-0610</td>
<td>1</td>
<td>CATCGCCCAATTCGCTATCTCACTACGAGGACTCGGCTCTCTGACTGCTCTTTA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>GGC-ATGAGGTGTCTGCTGTTCTGTTCTGAAGG</td>
<td>11</td>
</tr>
<tr>
<td>MAS94-0018</td>
<td>1</td>
<td>CATCGCCCAATTCGCTATCTCACTACGAGGACTCGGCTCTCTGACTGCTCTTTA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>GGC-ATGAGGTGTCTGCTGTTCTGTTCTGAAGG</td>
<td>11</td>
</tr>
<tr>
<td>MAS94-0241</td>
<td>1</td>
<td>CATCGCCCAATTCGCTATCTCACTACGAGGACTCGGCTCTCTGACTGCTCTTTA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>GGC-ATGAGGTGTCTGCTGTTCTGTTCTGAAGG</td>
<td>11</td>
</tr>
<tr>
<td>MAS94-0244</td>
<td>1</td>
<td>CATCGCCCAATTCGCTATCTCACTACGAGGACTCGGCTCTCTGACTGCTCTTTA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>GGC-ATGAGGTGTCTGCTGTTCTGTTCTGAAGG</td>
<td>11</td>
</tr>
<tr>
<td>MAS94-0351</td>
<td>1</td>
<td>CATCGCCCAATTCGCTATCTCACTACGAGGACTCGGCTCTCTGACTGCTCTTTA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>GGC-ATGAGGTGTCTGCTGTTCTGTTCTGAAGG</td>
<td>11</td>
</tr>
</tbody>
</table>

*Nucleotide positions of the analysed sequence correspond to those described by Edman & Kwon-Chung (1990). Only nucleotides 65–706 bp were unambiguously obtained for all alleles (see text for details). Underlined polymorphic nucleotide positions are located in introns. Nucleotide positions are presented vertically with the first polymorphic site at nucleotide position 67 and the last at 701.
**URA5 locus.** One diploid genotype (haplotype combination 2/12) was represented by four strains, and another one (8/11) was present in two strains. The other eight diploid genotypes were represented by one strain each (Table 2). The shared haplotype combination 2/12 was found in isolates from three different patients in the San Francisco area. The second shared haplotype combination (8/11) was found in two isolates from the same patient (MAS93-0315 and MAS93-0610). These strains might be considered identical except that they shared only one haplotype for the LAC gene (Fig. 1b, see also Xu et al., 2002).

**Phylogenetic analyses**

Of the 51 variable nucleotide sites, 43 had states shared by at least two haplotypes and were therefore phylogenetically informative. Eight sites were unique mutations possessed by only one haplotype each and these were phylogenetically uninformative. Fig. 1(a) presents one of the 12 MP trees among the 28 URA5 sequences from the 14 strains of serotype AD. Five representative sequences each of serotypes A and D were also included for haplotype identification and phylogenetic analysis. The MP trees had a length of 55 steps with a consistency index (CI) of 0.943, a

---

**Fig. 1.** (a) One of 12 MP trees for the 28 URA5 sequences from 14 serotype AD strains of C. neoformans. This tree has a consistency index of 0.943, a retention index of 0.993, and a total tree length of 56. For comparison, five representative sequences from serotype A (E1, CN-A, MMRL750, J10 and ZG280) and five from serotype D (B10, CN-D, J9, MMRL751 and MMRL757) were included in this figure. These ten sequences were reported in an earlier study (Xu et al., 2000b). Numbers above branches are bootstrap values > 50% and based on 1000 replicates. Designations for strains of serotypes A and D include the isolate name, geographical origin (CA, California; NYC, New York City, both in the USA) and serotype. For the 28 serotype AD sequences, designations correspond to those in Table 1 and are followed by ‘-1’ or ‘-2’ to indicate alleles within each strain (Table 2). Mid-point rooting is used for this phylogeny but the tree topology is identical to that when serotype B or C sequences were used as outgroups. The scale bar represents one nucleotide substitution. (b) One of 10 MP trees for the 28 LAC sequences from the same 14 serotype AD strains of C. neoformans. Genealogical relationships among these 28 sequences were reported in an earlier study (Fig. 1 in Xu et al., 2002). However, some of the reference strains are different. To be consistent with panel (a), the same ten reference strains of serotypes A and D were included in panel (b). Strain designations and figure labels are the same as those in (a).
five haplotypes were found for the LAC haplotypes within the serotype D group (Xu et al., 2000b). Each serotype AD strain contained two distinct alleles for both genes, with one allele identical or highly similar to those from strains of serotype D (allele 1) and the other to those from strains of serotype A (allele 2). These serotype-specific haplotype groups were used to analyse genealogical congruence between genes and to infer recombination within populations of serotypes A and D.

Despite the overall similarities between the genealogies from these two genes, there were several differences within each of the two haplotype groups. First, for the same 14 serotype AD strains, there were more URA5 haplotypes than LAC haplotypes within the serotype D group (Xu et al., 2002; Table 2; Fig. 1). Specifically, ten haplotypes were found among the 14 sequences for the URA5 gene but only five haplotypes were found for the LAC gene. However, the numbers of haplotypes clustered with serotype A strains were very similar between the two genes (five for LAC and four for URA5).

Second, based on phylogenetic patterns and bootstrap values (Fig. 1a), analysis of URA5 identified at least four hybridization events: (i) one generated strain MAS94-0351; (ii) one generated MAS93-0315 and MAS93-0610, two strains from the same patient; (iii) one generated strains MAS92-0022, MAS92-0793 and MAS92-0855, each from a different patient; and (iv) one generated the remaining eight hybrids. Similar analyses of the LAC gene identified at least three hybridization events (Fig. 1b, Xu et al., 2002): (i) one for strains MAS92-0793 and MAS92-0022; (ii) one for strains MAS92-0855 and MAS94-0351; and (iii) one for the remaining ten strains. The combined analyses would suggest at least five hybridization events: (i) one for strain MAS94-0351; (ii) one for strain MAS92-0855; (iii) one for strains MAS93-0315 and MAS93-0610; (iv) one for strains MAS92-0022 and MAS92-0793; and (v) one for the remaining eight strains. It should be noted that there are minor sequence differences among AD strains generated by each proposed hybridization event (Table 2).

Evidence for recombination within both serotype A and serotype D haplotype groups

Phylogenetic analyses identified gene genealogy incongruence between the two genes within both A and D haplotype groups (Fig. 1). To determine the statistical significance of the observed incongruence, a partition homogeneity (PH) test was performed for the two haplotype groups: one group included all the 14 allele 1's of both genes from each serotype AD strain (i.e. those that clustered with the serotype D sequences) and the other included the 14 allele 2's of both genes from each serotype AD strain (i.e. those that clustered with the serotype A sequences). The PH test of each group showed that the two genealogies were significantly incongruent for both A and D serotype clusters (ILD < 0.05 for both groups). This result is consistent with recombination within each of the two serotypes. Visual inspection determined that the genealogical incongruence was caused by the following strains: MAS94-0351, MAS93-0315, MAS93-0610 and MAS92-0855. When these strains were excluded from the PH tests, the two genealogies were congruent (ILD > 0.2).

DISCUSSION

Similar to other studies of genetic variation within serotype AD strains and between serotype AD and other serotypes of C. neoformans (e.g. Brandt et al., 1995, 1996; Boekhout et al., 2001; Xu et al., 2000a, b), the URA5 sequences provide abundant evidence for high genetic variability within and among strains of serotype AD. Based on sequence from 642 nucleotides of URA5, 14 unique haplotypes were identified among the 28 sequenced alleles. Analyses of the URA5 gene confirmed that strains of serotype AD resulted from recent hybridizations between strains of serotypes A and D, as recently reported (Xu et al., 2002). Aside from the three hybridization events revealed by the LAC gene genealogy (Fig. 1b, Xu et al., 2002), the combined analyses of URA5 and LAC genes identified at least two additional hybridization events (see Results and Fig. 1). More importantly, the combined genealogical analysis revealed recombination within both serotype A and serotype D haplotype groups.

Previous population studies indicated a predominantly clonal population structure of C. neoformans. Evidence of a clonal population structure included: (i) over-representation of certain genotypes; (ii) lack of segregation among alleles at different loci; and (iii) overall congruence of gene genealogies (e.g. Brandt et al., 1996; Xu et al., 2000b). However, gene genealogical analyses in our earlier study also revealed hybridization and recombination among serotypes (Xu et al., 2000b). Because those earlier samples were from geographically diverse areas around the globe and included only one or two strains of each serotype from each area, the question of whether recombination occurred within a serotype could not be addressed (Xu et al., 2000b). Though the possibility of recombination within serotypes has been suggested [Franzot et al., 1999; Taylor et al. (1999)] based on data from Brandt et al. (1996); Xu et al. (2000b)], the evidence was largely circumstantial and inconclusive.

This study differs from previous reports on natural populations of serotypes A and D. First, the 14 strains analysed here are unique. Previous population studies of serotypes A and D focused on strains that were serotype A or D, but not serotype AD. Intuitively, that was logical.
However, natural samples of serotype A or serotype D are often greatly skewed, favouring MATα. Such biased mating type ratios could manifestly contribute to the appearance of a clonal population structure. In contrast, the serotype AD strains used here have a balanced mating type ratio, with most strains having both MATα and MATα (Table 1). The chances of detecting recombination should be greater in samples with comparable numbers of each mating type. Second, 12 of these 14 serotype AD strains were from one geographical region (the San Francisco area; the other two strains were one each from Georgia and Texas). Although we have little knowledge of where or when the patients initially acquired their infections, the possibility of detecting recombination is probably higher among strains from the same geographical area, where there could be the opportunity for contact and sexual recombination. Third, the two genes analysed here are highly polymorphic. Highly polymorphic gene sequences permit the construction of robust phylogenies, leading to unambiguous inferences regarding hybridization and recombination.

Our studies showed that combined analyses of the two genes identified more hybridization events than those based on each individual gene. Analyses of additional genes might reveal more hybridization and recombination events. Recent studies using gene genealogical comparison to infer recombination and speciation have typically used four or more genes (for a review see Taylor et al., 1999). Indeed, our earlier study used four genes located at different parts of the genome [LAC, URA5, the internal transcribed spacer region of rRNA (ITS), and the mitochondrial large ribosomal gene (mtLrRNA)]. However, neither the ITS nor the mtLrRNA gene showed phylogenetically informative variation among geographically diverse strains within serotypes A and D (Xu et al., 2000b). We have also screened portions of other genes using PCR-RFLP, including the nuclear genes GPA1 and ADE2 and the mitochondrial gene NADH dehydrogenase subunit 2. Unfortunately, no phylogenetically informative variation among strains within either serotype A or D was found (unpublished data; Xu, 2002). Nevertheless, using the two highly polymorphic loci, our analyses clearly identified multiple hybridizations between strains of serotypes A and D and recombination within populations of both serotypes A and D.

While our analyses of URA5 and LAC genes using the 14 serotype AD strains provided unambiguous evidence for recombination within populations of both serotypes A and D, we would like to stress that our data are insufficient for estimating the frequencies of recombination in environmental populations of these two serotypes. To obtain such estimates, larger sample sizes are needed. In addition, increasing the number of genes could also enhance the probability of detecting recombination events. Due to differences in the geographical distribution of these serotypes, it is likely that populations of C. neoformans from different locations will show different degrees of clonality and recombination.

**Implications for environmental populations of C. neoformans**

Our inferred population structures of serotypes A and D derived from clinical strains of serotype AD are different from direct analyses of strains of serotypes A and D from clinical samples. Based on our analyses, we believe serotype AD strains are more reflective of environmental serotype A and D populations than of clinical serotype A and D populations. For example, one of the consistent observations of clinical samples is that strains of MATα predominate. Indeed, no MATα strain was found among strains of serotypes A (324 strains), B (3 strains) and D (12 strains) isolated from four different geographical areas in the USA (San Francisco, Georgia, Texas and Alabama). All 339 strains had the MATα allele (Yan et al., 2002). However, analyses of mating types from strains of serotype AD clearly suggested that both mating types existed in this sample. Among the 14 strains of serotype AD analysed here, nine contained the MATα allele from serotype A (Aa), three contained the MATα allele from serotype A (Az), and the remaining two contained neither mating type allele from serotype A (Table 1, Yan et al., 2002). Similarly, ten strains had the Dx allele, three strains had the Da allele and one contained neither mating type allele from serotype D. Though the mating type allele ratios were not 1:1 among the 14 strains for either serotype A or D, they were much more balanced than in typically analysed clinical populations of serotypes A or D. Furthermore, because these serotype AD strains arose relatively recently as a result of hybridization between strains of serotypes A and D (Xu et al., 2002), strains of serotype A with MATα must exist in the environment in North America at present or have existed in the recent past. Strains of serotype A, MATα were until very recently thought to be extinct (Lengeler et al., 2000). Our results here suggest that these strains not only exist in North America, but that they are recombining with strains of serotype A and MATα. Coupled with evidence for recent hybridization between strains of serotypes A and D, the recombining populations of both serotypes A and D indicate continuously evolving natural populations of this important human pathogenic fungus.

**ACKNOWLEDGEMENTS**

We thank Dr Mary Brandt for strains and Guizhen Luo for technical assistance in the early stages of this study. Research support for this project was provided by grants to J. Xu from McMaster University, the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Premier’s Research Excellence Award (PREA) of Ontario, Genome Canada, the Canadian Foundation for Innovation (CFI) and the Ontario Innovation Trust (OIT), and by grants to T. G. Mitchell from the US Public Health Service National Institutes of Health (AI 25783 and AI 44975). All 14 serotype AD strains analysed in this study were collected through the CDC Cryptococcal Active Surveillance program.


