Multiple origins of hybrid strains of *Cryptococcus neoformans* with serotype AD

Jianping Xu,¹ Guizhen Luo,² Rytas J. Vilgalys,³ Mary E. Brandt⁴ and Thomas G. Mitchell²

Author for correspondence: Jianping Xu. Tel: +1 905 525 9140 ext. 27934. Fax: +1 905 522 6066.
e-mail: jpxu@mcmaster.ca

¹ Department of Biology,
McMaster University,
Hamilton, Ontario,
L8S 4K1, Canada

² Department of
Microbiology, Duke
University Medical
Centre, Durham,
NC 27710, USA

³ Department of Biology,
Duke University, Durham,
NC 27706, USA

⁴ Mycotic Diseases Branch,
Centers for Disease
Control and Prevention,
Atlanta, USA

***Cryptococcus neoformans* is a major pathogen of humans throughout the world. Using commercial mAbs to capsular epitopes, strains of *C. neoformans* manifest five distinct serotypes – A, B, C, D and AD. Previous studies demonstrated significant divergence among serotypes A, B, C and D, which are thought to be haploid. In this study the origins and evolution of strains of serotype AD were investigated. A portion (537 bp) of the laccase gene was cloned and sequenced from 14 strains of serotype AD. Each strain contained two different alleles and sequences for both alleles were obtained. These sequences were compared to those from serotypes A, B, C and D. This analysis indicated that each of the 14 serotype AD strains contained two phylogenetically distinct haplotypes: one haplotype was highly similar to the serotype A group and the other to the serotype D group. To explain the origins of these serotype AD strains, genealogical analysis is consistent with at least three recent and independent hybridization events. The results demonstrate that the evolution of *C. neoformans* is continuing and dynamic.**

Keywords: laccase, gene genealogy, multiple hybridizations

INTRODUCTION

Cryptococcus neoformans is an important fungal pathogen of humans and other mammals throughout the world. It is a basidiomycetous yeast, typically with a haploid nucleus (Casadevall & Perfect, 1998). From the 1980s to the early 1990s, due to the high number of immunocompromised hosts, the incidence of cryptococcal disease increased dramatically (Casadevall & Perfect, 1998; Mitchell & Perfect, 1995). However, the incidence has been decreasing since the mid-1990s in the United States (Ponce de Leon *et al.*, 1999) and Europe (Kwon-Chung *et al.*, 2000), perhaps because of improved therapies for HIV. Elsewhere, particularly in Africa, the incidence of cryptococcosis remains high and is expected to continue to rise (Kwon-Chung *et al.*, 2000).

Using commercial mAbs to capsular epitopes, strains of *C. neoformans* manifest five distinct serotypes – A, B, C, D and AD (Kabasawa *et al.*, 1991). A small number of strains do not react with any of the four serotype-specific antibodies. Recent studies show that serotypes

often show significant divergence at the molecular level (e.g. Xu *et al.*, 2000c). Three varieties have been proposed in recognition of the profound divergence within this biological species: *C. neoformans* var. *neoformans* represents serotype D strains, *C. neoformans* var. *grubii* represents serotype A strains and *C. neoformans* var. *gattii* represents serotypes B and C strains (Kwon-Chung *et al.*, 1982; Franzot *et al.*, 1999). However, this three-variety classification system ignores the varietal status of serotype AD strains. More 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 *C. neoformans* (Sanfelice) Vuillemin and strains of serotypes B and C in a new species, *Cryptococcus bacillisporus* Kwon-Chung. This nomenclature reverts to the original classification proposed for the teleomorphic species of *Filobasidiella* (Kwon-Chung, 1976; Aulakh *et al.*, 1981).

While current studies recognize that serotype AD strains are potentially different from strains of other serotypes, there is little consensus about their origin and molecular divergence from other serotypes. Though clinical isolates of serotype AD are not as common as serotype A

Abbreviations: K2P, Kimura two-parameter; MLEE, multilocus enzyme electrophoresis; 6PD, 6-phosphoglucanate dehydrogenase.

strains, they can be comparable to or occur in higher frequency than other serotypes in certain geographic areas. For example, in one survey, 8 of 13 (over 60%) strains obtained from Belgium were serotype AD (Kwon-Chung & Bennett, 1984). Similarly, 14% of isolates found from 1992 to 1994 in San Francisco (CA, USA) were serotype AD, which exceeded the combined number of isolates of serotypes B, C and D from this region (Brandt *et al.*, 1996). Aside from the classification dilemma posed by serotype AD strains, the potentially more important issue is their impact on the continuing evolution of *C. neoformans*. This study investigates the origin and evolution of strains of serotype AD.

C. neoformans is a heterothallic species with two alternative mating types, *a* and α . Under suitable conditions, yeast cells with opposite mating types can fuse to form dikaryotic hyphae. In the terminal cell (termed basidium) of a dikaryotic hypha, nuclear fusion and meiosis occur to produce four chains of haploid basidiospores (Kwon-Chung, 1976). When strains of serotypes A and D mate in the laboratory, meiotic progenies often contain alleles from both parental strains and are diploid or aneuploid (e.g. Lengeler *et al.*, 2001). These laboratory observations coupled with genotypic analysis of clinical serotype AD strains (Brandt *et al.*, 1993, 1995; Boekhout *et al.*, 2001) suggest that strains of serotype AD are likely to be genetic hybrids. As in plants and animals, we use the terms 'hybridization' and 'hybrid' to denote the process or a product of the process in which an offspring is generated from a mating between two genetically divergent parental strains (e.g. strains of different races, breeds, varieties, subspecies, species or genera). Strains of serotypes A and D belong to two different varieties and they have diverged from each other for millions of years (Xu *et al.*, 2000c). Therefore, mating between strains of serotypes A and D is referred to as hybridization. Hybridization differs from recombination. Recombination is generally defined as the formation of new combinations of genes in progeny that did not occur in parents, by the processes of crossing-over and independent assortment during meiosis.

Though the hypothesis of hybrid origin has been suggested for serotype AD strains, an alternative hypothesis of an ancient single origin has not been rejected. In addition, variations of the hybrid origin hypothesis have not been vigorously tested. Here, we propose and test four possible hypotheses about the origin and evolution of strains of serotype AD.

The first hypothesis (hypothesis I) is that serotype AD is ancient and its patterns of genetic variation are similar to other serotypes. Under this hypothesis, serotype AD and other serotypes have undergone similar molecular divergence and serotype AD strains should form a distinct cluster comparable to, but separate from the other serotypes.

The second hypothesis (hypothesis II) is that serotype AD strains have a single ancient hybrid origin resulting from hybridization between strains of serotypes A and

D. In this case, the observed genetic heterogeneity among serotype AD strains would be attributable to the subsequent accumulation of mutations. This hypothesis predicts that because mutations have accumulated over millions of years, no identical sequences should be found between strains of serotype AD and strains of serotypes A or D.

The third hypothesis (hypothesis III) is that serotype AD strains have a single recent origin, resulting from hybridization between strains of serotypes A and D, and that the widespread geographic distribution and extensive genetic heterogeneity of serotype AD strains are due to recent dispersal and genomic changes. Under this hypothesis, all serotype AD sequences should coalesce to two points: one a cluster within serotype A, and the other, a cluster within serotype D.

The fourth hypothesis (hypothesis IV) proposes multiple hybridization events ranging from ancient to recent times. This hypothesis predicts that sequences from strains of serotype AD will show significant heterogeneity and divergence as a result of differences among the original hybridizing parental strains and the accumulation of mutations following the hybridization events.

Current evidence supports one or more of the hybrid origin hypotheses (II, III and IV). First, strains of serotype A isolated from either nature or patients can mate with serotype D strains (e.g. Xu *et al.*, 2000a). Second, clinical isolates of serotype AD typically exhibit heterozygosity at several enzymic loci, sharing alleles from both serotypes A and D strains (Brandt *et al.*, 1993, 1995). Third, DNA fingerprints from serotype AD strains reveal patterns of genetic similarity that are intermediate between strains of serotypes A and D (e.g. Boekhout *et al.*, 2001). Fourth, a study by Lengeler *et al.* (2001) showed that some serotype AD strains are diploid (2N) or aneuploid and possess both serotype A- and D-specific alleles at several loci. In contrast, strains of serotypes A or D were haploid and contained only one allele for each single-copy gene (Brandt *et al.*, 1993, 1995; Xu *et al.*, 2000c). However, none of these studies were able to refute hypothesis I or to distinguish among hypotheses II, III and IV.

We have used a gene genealogical approach to evaluate the various hypotheses about the origin and evolution of serotype AD strains in *C. neoformans*. The 14 serotype AD strains examined here were from three locations in the US and obtained during the 1992–1994 *Cryptococcus* surveillance conducted by the Centers for Disease Control and Prevention in Atlanta, GA, USA. In a previous study, we screened four genes in 34 strains of *C. neoformans* and found the laccase (*LAC*) gene to be the most polymorphic and phylogenetically informative (Xu *et al.*, 2000c). Therefore, we have used the same fragment of the *LAC* gene to test the four possible hypotheses. In addition, based on the divergence time among sequences of 18.5 million years between serotypes A and D (Xu *et al.*, 2000c), we estimated the time at which these hybridization events occurred.

METHODS

Strains. The 14 isolates analysed in this study were collected from three geographic locations 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 strains were chosen to represent all three multilocus enzyme electrophoresis (MLEE) genotypes discovered in that surveillance for serotype AD strains (Brandt *et al.*, 1993, 1995). Table 1 presents their geographic origins, site and year of isolation and the MLEE genotypes determined in an earlier study (Brandt *et al.*, 1995). Two pairs of strains were isolated from the same two patients. Strains MAS93-0315 and MAS93-0610 were isolated from the same body site (cerebrospinal fluid) of one patient and strains MAS94-0241 and MAS94-0244 were isolated from different body sites of another patient. Both patients were from San Francisco, CA, USA. Each of the other strains was isolated from a different patient (Table 1).

DNA manipulations. DNA was isolated from each strain as described previously (Xu *et al.*, 2000b). A portion of the diphenol oxidase gene (laccase or *LAC*) of each strain was amplified, sequenced and analysed. The fragment was amplified using the following oligonucleotide primers: forward, 5'-GGCGATACTATTATCGTA-3'; reverse, 5'-TTCTGGAG-TGGCTAGAGC-3'. The amplified DNA fragment corresponded to nt 401–987 as reported for a serotype D strain in the study by Williamson (1994; GenBank accession no. L22866). However, because ends of sequences were usually not clear on typical sequencing gels, only the unambiguous nucleotides from position 429–962 (a total of 534 nt) were analysed for all strains in this study. This fragment contains four introns, three complete exons and partial sequences from two other exons (see Williamson, 1994). Because of insertions and deletions, the completely aligned sequences stretched to a total 537 nt (see below).

A typical PCR reaction contained 10 µl (~1 ng) diluted genomic DNA template, 0.5 units Amplitaq DNA polymerase, 0.2 µM each primer and 0.2 mM each dNTP 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 finally 7 min extension at 72 °C. PCR products were cleaned by using Wizard spin columns (Promega) and sequenced using an Applied Biosystems PRISM 373 (ABI 373) automated sequencer with dRhodamine-labelled terminators (PE Applied Biosystems) following the manufacturer's instructions. However, we were unable to obtain good *LAC* sequences from any of the 14 strains by the direct PCR sequencing method. This result suggested sequence heterogeneity at the *LAC* 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 product into *Escherichia coli* following the manufacturer's instructions. For each of the 14 strains, 10 random *E. coli* colonies were picked, amplified with the *LAC* primers and digested with restriction enzyme *DpnI* to screen for different alleles. (Here we use the term 'allele' to refer to different sequences for a locus within a strain. The term 'haplotype' will be used for distinct sequences in the whole sample.) For each of the two strains MAS93-0315 and MAS93-0610, two clones representing each of two *DpnI* restriction digest patterns were sequenced (four clones total for each strain). For each of the remaining 12 strains, only one clone of each *DpnI* restriction digest pattern was sequenced. Two alleles were found in each of the 14 strains. A representative of each allele was sequenced for each strain using the ABI 373. These sequences exhibited no ambiguity. Sequences from both strands were generated, aligned using Sequencher 3.1.1 (Gene Codes) and optimized visually.

Data analyses. Phylogenetic analysis was performed with PAUP 4.0 (Swofford, 2001). Maximum-parsimony trees were identified using heuristic searches based on 500 random sequence additions (Swofford, 2001). 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 identifies phylogenetically distinct and statistically well supported sequence clusters. Only statistically robust sequence clusters can be used to infer hybridization events (see Results below).

To test the various hypotheses about the origin and evolution of serotype AD strains, we also included the published *LAC*

Table 1. Serotype AD strains of *C. neoformans* investigated in this study

Strain	Geographic origin	MLEE-type*	Source†	Year
MAS92-0022	San Francisco	7	CSF	1992
MAS92-0086	San Francisco	21	Blood	1992
MAS92-0153	San Francisco	21	Blood	1992
MAS92-0189	Georgia	21	CSF	1992
MAS92-0224	San Francisco	7	CSF	1992
MAS92-0668	San Francisco	21	CSF	1992
MAS92-0793	San Francisco	21	CSF	1992
MAS92-0855	San Francisco	7	CSF	1992
MAS93-0315	San Francisco	5	CSF	1993
MAS93-0610	San Francisco	5	CSF	1993
MAS94-0018	San Francisco	21	Peritoneal fluid	1994
MAS94-0241	San Francisco	7	Bronchial wash	1994
MAS94-0244	San Francisco	7	Prostate	1994
MAS94-0351	Texas	21	CSF	1994

* MLEE genotypes correspond to those in Brandt *et al.* (1993, 1995).

† CSF, Cerebrospinal fluid.

sequences from 13 serotype A strains, 6 serotype D strains, 2 serotype B strains and 5 serotype C strains (Xu *et al.*, 2000c). For strains of serotypes A, B, C and D, direct sequencing using PCR products without cloning showed no sequence ambiguity and each strain had only one allele at any of the four genes analysed, including the *LAC* gene (Xu *et al.*, 2000c).

Predictions from various hypotheses were compared to those observed in the molecular sequence data. To assess the amount of sequence variation within each serotype group, we calculated the Kimura two-parameter (K2P) distance (Kimura, 1980) between all pairs of sequences within each of the five serotypes A, B, C, D and AD (sequence data for serotypes A, B, C and D were from Xu *et al.*, 2000c). Sequence divergence between serotype A and D groups was also calculated. In our previous study, ambiguous alignment was common between sequences from serotypes B and C and those from serotypes A and D; therefore only unambiguous alignments were analysed in that study. In this study, because we were only analysing sequences from serotypes A, D and AD, there was little ambiguity in the alignments. Therefore, all nucleotides were used for analysis. Student's *t*-test was used to compare genetic differences within and between groups.

Significantly higher sequence diversity among serotype AD sequences compared with other serotypes and the close relatedness between serotype AD sequences and those from other serotypes were inconsistent with the hypothesis of an ancient single origin of serotype AD strains (hypothesis I). A single ancient hybrid origin of serotype AD strains would predict that no sequences would be identical between serotype AD strains and those from either serotype A or D strains. The observation of such identical sequences would reject hypothesis II.

To test whether serotype AD strains derived from a single recent hybridization between a serotype A strain and a serotype D strain, we obtained the 'within serotype' sequence diversity for serotype A and D groups (see above). Similar measures were obtained for each of the two allelic classes in the serotype AD group (see Results below): one was the serotype 'A allele' class and the other was the serotype 'D allele' class. A recent single hybridization would predict that sequences within each of the two allelic classes should have very low sequence diversity compared to the entire serotype A and D groups. Furthermore, under this hypothesis, all the serotype 'A alleles' from serotype AD strains should form one tight cluster within the serotype A lineage and all the serotype 'D alleles' from serotype AD strains would form another distinct clade within the serotype D lineage. Lack of support for these predictions would reject hypothesis III.

In contrast, under hypothesis IV, we should expect the following results. First, we should observe two highly divergent clusters of sequences, one similar to serotype A sequences and the other similar to serotype D sequences. Second, within each of the two serotype AD sequence clusters, we should observe sequence diversities similar to those of serotypes A or D. Third, statistically robust sequence clusters should allow the identification of distinct independent hybridization events. Furthermore, if hybridization events were recent, some sequences from serotype AD strains are likely to be identical to those from serotype A or D strains.

Assuming hybrid origins of the serotype AD strains, we can estimate the time of hybridization. To obtain such estimates, we first tested whether the *LAC* gene evolved in a molecular clock fashion. K2P distances were calculated from the maximum-likelihood tree (Felsenstein, 1981) based on empirical nucleotide frequencies of a selected subset of DNA

sequences (see Fig. 1 for sequences tested). Maximum-likelihood estimates of the most parsimonious trees with and without a molecular clock demonstrated that the *LAC* genes evolved in a clock-like manner ($P > 0.2$). Estimates of the time when hybridization occurred in each putative hybrid were based on the maximum-likelihood trees and assumed a divergence time of 18.5 million years between serotypes A and D (Xu *et al.*, 2000c). K2P distances were calculated between each of the two sequences in each serotype AD strain and its closest serotype A or D sequence on the phylogenetic tree.

RESULTS

Sequence variation at the *LAC* gene within and among strains

Of the 537 aligned nucleotides, a total of 63 were variable among the 28 DNA fragments from the 14 serotype AD strains (Table 2). Only the 63 polymorphic nucleotide sites are shown in Table 2. The remaining 474 sites were identical among all 28 sequences and therefore offered no discriminating power for haplotype identification. Among the 63 polymorphic sites, 36 were located in introns (see underlined nucleotide positions in Table 2) and 27 in exons. The four introns included a total of 226 aligned nucleotide positions and the exons had 311 nt. Intron nomenclature for the *LAC* gene as described by Williamson (1994) corresponds to the following nucleotide positions in this study: intron IV, 29–96; intron V, 159–210; intron VI, 327–380; and intron VII, 465–516. Overall, introns had significantly higher proportions of polymorphic sites than exons (χ^2 value = 5.95, degrees of freedom = 1, $P < 0.05$).

Among these 28 sequences, 10 unique haplotypes were found (Table 2). These haplotypes are present in different frequencies, with haplotype 1 found in nine strains, haplotype 4 in two strains, haplotype 6 in four strains and haplotype 7 in seven strains. Haplotypes 2, 3, 5, 8, 9 and 10 were found in one strain each. No two sequences from the same strain are identical at the sequenced *LAC* region (Table 2).

A total of nine diploid genotypes were found among the 14 strains at the *LAC* locus. One genotype (haplotype combination 1/7) was represented by four strains. Two genotypes (haplotype combinations 1/6 and 4/6) were represented by two strains each. The other six genotypes were represented by one strain each (Table 2). The shared haplotype combination 1/6 was found from different patients in the same geographic area (San Francisco), both obtained in 1992. The haplotype combination, 4/6, was found from two different patients at different geographic areas and at different times: strain MAS92-0855 was obtained in 1992 from San Francisco and strain MAS94-0351 in 1994 from Texas.

Two strains, MAS93-0315 and MAS93-0610, from one patient shared one common haplotype (haplotype 1), but their second haplotypes (haplotypes 7 and 10; Table 2) differed by one substitution at position 203. Similarly, the other pair of strains, MAS94-0241 and MAS94-0244, from a different patient also shared one common

Table 2. Polymorphic nucleotide positions in a portion of the laccase gene from the 14 serotype AD strains of *C. neoformans* analysed in this study

Underlined nucleotide positions are located in introns. Intron nomenclature corresponds to that in Williamson (1994).

Strain	Allele	Polymorphic nucleotide position				Haplotype
		Intron IV	Intron V	Intron VI	Intron VII	
		00000000000000111111112222222222333333333333444444444444555555				
		012345556678881567899900024467893445577889991122336778889000233				
		814312792634899593612736794675544460405280371456283483455678336				
MAS92-0022	1	GTCCCCTAGC-CTAGTCATTTGTGTGATCACCCC-GCATCGGCCTGACTCATTCAA-GACTAA				1
	2	GGATGC-GATTC--TATGC--TCCAACAATTTTTGAGGCTACTCCAACCTGATTGGAGGTCGG				6
MAS92-0086	1	ATCCCCTAGC-CTAGTCATTTGTGTGATCACCCC-GCATCGGCCTGACTCATTCAA-GACTAA				2
	2	GGATGC-GATTT--TATGC--TCCAACAATTTTTGAGGCCACCCCAACCTGATTGGAAGTCGG				7
MAS92-0153	1	GTCCCCTAGC-CTAGTCATTTGTGTGATCACCCC-GCATCGGCCTGACTCATTCAA-GACTAA				1
	2	GGATGC-GATTT--TATGC--TCCAACAATTTTTGAGGCCACCCCAACCTGACTGGAAGTCGG				8
MAS92-0189	1	GTCCCCTAGC-CTAGTCATTTGTGTGATCACCCC-GCATCGGCCTGACTCATTCAA-GACTAA				1
	2	GGATGC-GATTT--TATGC--TCCAACAATTTTTGAGGCCACCCCAACCTGATTGGAAGTCGG				7
MAS92-0224	1	GTCCCCTAGC-CTAGTCATTTGTGTGATCACCCC-GCATCGGCCTGACTCATTCAA-GACTAA				1
	2	GGATGC-GATTT--TATGC--TCCAACAATTTTTGAGGCCACCCCAACCTGATTGGAAGTCGG				7
MAS92-0668	1	GTCCCCTAGC-CTAGTCATTTGTGTGATCACCCC-GCATCGGCCTGGCTCATTCAA-GACTAA				3
	2	GGATGC-GATTT--TATGC--TCCAACAATTTTTGAGGCCACCCCAACCTGATTGGAAGTCGG				7
MAS92-0793	1	GTCCCCTAGC-CTAGTCATTTGTGTGATCACCCC-GCATCGGCCTGACTCATTCAA-GACTAA				1
	2	GGATGC-GATTC--TATGC--TCCAACAATTTTTGAGGCTACTCCAACCTGATTGGAGGTCGG				6
MAS92-0855	1	GTCCCTTAGC-CTAGTCATTTGTGTGATCACCCC-GCATCGGCGTGATTCATTCAA-GACTAA				4
	2	GGATGC-GATTC--TATGC--TCCAACAATTTTTGAGGCTACTCCAACCTGATTGGAGGTCGG				6
MAS93-0315	1	GTCCCCTAGC-CTAGTCATTTGTGTGATCACCCC-GCATCGGCCTGACTCATTCAA-GACTAA				1
	2	GGATGC-GATTT--TATGC--TTCAACAATTTTTGAGGCCACCCCAACCTGATTGGAAGTCGG				10
MAS93-0610	1	GTCCCCTAGC-CTAGTCATTTGTGTGATCACCCC-GCATCGGCCTGACTCATTCAA-GACTAA				1
	2	GGATGC-GATTT--TATGC--TCCAACAATTTTTGAGGCCACCCCAACCTGATTGGAAGTCGG				7
MAS94-0018	1	GTCCCCTAGC-CTAGTCATTTGTGTGATCACCCC-GCATCGGCCTGACTCATTCAA-GACTAA				1
	2	GGGTGC-GATTT--TATGC--TCCAACAATTTTTGAGGCCACCCCAACCTGATTGGAAGTCGG				9
MAS94-0241	1	GTCCTCCTAGC-CTAGTCATTTGTGTGATCACCCC-GCATCGGCCTGACTCATTCAA-GACTAA				5
	2	GGATGC-GATTT--TATGC--TCCAACAATTTTTGAGGCCACCCCAACCTGATTGGAAGTCGG				7
MAS94-0244	1	GTCCCCTAGC-CTAGTCATTTGTGTGATCACCCC-GCATCGGCCTGACTCATTCAA-GACTAA				1
	2	GGATGC-GATTT--TATGC--TCCAACAATTTTTGAGGCCACCCCAACCTGATTGGAAGTCGG				7
MAS94-0351	1	GTCCCTTAGC-CTAGTCATTTGTGTGATCACCCC-GCATCGGCGTGATTCATTCAA-GACTAA				4
	2	GGATGC-GATTC--TATGC--TCCAACAATTTTTGAGGCTACTCCAACCTGATTGGAGGTCGG				6

haplotype (haplotype 7) and their second haplotypes (haplotypes 1 and 5) differed by 1 nt at position 33. For several reasons, these results reflect true genetic difference and not experimental error. First, PCR error should result in multiple peaks at individual sites and therefore ambiguous base calling. However, all our 537 sites were unambiguous for all 28 sequences. Second, for the two strains where four clones were taken from each (see Methods and Results), only two haplotypes were found from each strain, suggesting PCR and cloning errors were very low or negligible. Third, the recovery of identical haplotypes from different strains further suggests that experimental error was minimal and our protocols were robust.

Fig. 1 presents the phylogenetic relationships among the 28 sequences from the 14 strains of serotype AD. Phylogenetic evidence indicated that each strain has two distinct alleles, with one clustered with serotype A

strains and the other with serotype D strains. For ease of comparison, only four reference sequences from serotype A (E1, CN-A, M0013 and J10) and three from serotype D (B10, CN-D and MMRL751) were included in Fig. 1. These seven sequences were chosen because of their high sequence similarities to the 28 serotype AD sequences. Our earlier analysis indicated that geographic location did not affect sequence diversity in *C. neoformans* (Xu *et al.*, 2000c). However, in the statistical analysis of sequence diversity, all 13 serotype A and 6 serotype D sequences were included (see below and Table 3).

Rejection of hypothesis I

The mean K2P distances between pairs of sequences are presented in Table 3. Sequence diversity within the 14 serotype AD strains is significantly higher than those

Fig. 1. One of 10 most parsimonious trees for the 28 LAC sequences from 14 serotype AD strains of *C. neoformans*. This tree has a consistency index of 0.957, a retention index of 0.996 and a total tree length of 69. For ease of comparison, only four reference sequences from serotype A (E1, CN-A, M0013 and J10) and three from serotype D (B10, CN-D and MMRL751) were included. Among the 13 serotype A sequences and 6 serotype D sequences analysed in an earlier study (Xu *et al.*, 2000c), these seven sequences showed the highest sequence similarities to the 28 serotype AD sequences. Numbers above each branch are bootstrap values > 50% based on 1000 replicates. Strain designations for the serotype A and D strains indicate the isolate name, geographic origin (CA, California; NYC, New York City; NC, North Carolina) and serotype. For the 28 serotype AD sequences, strain 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 an outgroup. Scale bar represents 1 nt substitution.

Table 3. K2P distances between pairs of alleles within and between groups of serotypes

Groups	No. of pairwise comparisons	Mean K2P distance ($\pm$ SD)
1. Within serotype A	91	0.00525 ($\pm$ 0.00475)
2. Within serotype D	21	0.00304 ($\pm$ 0.00218)
3. Between serotypes A and D	98	0.09739 ($\pm$ 0.00267)
4. Within serotype B	1	0.01183
5. Within serotype C	10	0.04558 ($\pm$ 0.02699)
6. Within the 14 serotype AD strains	378	0.06269 ($\pm$ 0.05645)
7. Within allele group 2 (serotype A group) of the 14 serotype AD strains	91	0.00589 ($\pm$ 0.00565)
8. Within allele group 1 (serotype D group) of the 14 serotype AD strains	91	0.00282 ($\pm$ 0.00335)
9. Between alleles 1 and 2 of the 14 serotype AD strains	196	0.11687 ($\pm$ 0.00364)

from within either serotypes A or D (Table 3; over 10-fold difference; compare K2P distance on line 6 with K2P distances on lines 1 and 2; $P < 0.001$). In addition,

phylogenetic evidence reveals the close relatedness between individual haplotypes from serotype AD strains and those from serotype A or D strains (Fig. 1). These

Table 4. Estimates for the age of hybridization based on laccase genealogy

Strain	Allele	Haplotype*	Node†	K2P coalescent distance‡	$10^{-6} \times$ Time of hybridization (years)§
MAS92-0022	1	1	d	0	0
MAS92-0022	2	6	h	0	
MAS92-0086	1	2	b	0.001878	0.297–2.158
MAS92-0086	2	7	g	0.013642	
MAS92-0153	1	1	d	0	0–2.142
MAS92-0153	2	8	g	0.013543	
MAS92-0189	1	1	d	0	0–2.155
MAS92-0189	2	7	g	0.013642	
MAS92-0224	1	1	d	0	0–2.155
MAS92-0224	2	7	g	0.013642	
MAS92-0668	1	3	c	0.001877	0.297–2.155
MAS92-0668	2	7	g	0.013642	
MAS92-0793	1	1	d	0	0
MAS92-0793	2	6	h	0	
MAS92-0855	1	4	f	0.003782	0–0.598
MAS92-0855	2	6	h	0	
MAS93-0315	1	1	d	0	0–2.135
MAS93-0315	2	10	g	0.013493	
MAS93-0610	1	1	d	0	0–2.155
MAS93-0610	2	7	g	0.013642	
MAS94-0018	1	1	d	0	0–2.142
MAS94-0018	2	9	g	0.013543	
MAS94-0241	1	5	e	0.001878	0.297–2.155
MAS94-0241	2	7	g	0.013642	
MAS94-0244	1	1	d	0	0–2.155
MAS94-0244	2	7	g	0.013642	
MAS94-0351	1	4	e	0.003782	0–0.598
MAS94-0351	2	6	h	0	

* Haplotypes correspond to those in Table 2.

† Nodes correspond to those in Fig. 1.

‡ Values of K2P distances were between serotype AD sequences and their closest serotype A or D sequences as shown in Fig. 1, indicated by the corresponding nodes.

§ Xu *et al.* (2000c) was used as a reference for the time range of hybridization. For each strain, estimates were obtained from both alleles.

results therefore are inconsistent with hypothesis I of a single ancient origin of serotype AD strains.

Rejection of hypothesis II

Of the 28 serotype AD sequences, 13 were identical to either serotype A (haplotype 6) or serotype D (haplotype 1) recovered from North America (Fig. 1). Such a high frequency of sequence identity is inconsistent with hypothesis II of a single ancient hybrid origin for serotype AD strains.

Rejection of hypothesis III

The third hypothesis is that there is a single recent hybrid origin of serotype AD and that the wide distribution of this serotype is the result of recent dispersal of this hybrid. This hypothesis is also rejected

for the following reasons. First, each of the two allelic classes in the 14 serotype AD strains exhibited high sequence diversity, comparable to those within serotypes A and D, respectively (Table 3; compare K2P distances on lines 1 and 7, and lines 2 and 8; $P > 0.9$ in both comparisons). Second, phylogenetic evidence revealed that only two common haplotypes (1 and 6) were shared between the serotype AD strains and those from either the serotype A strains (haplotype 6) or the serotype D strains (haplotype 1) (Fig. 1 and Xu *et al.*, 2000c). The other eight haplotypes in serotype AD strains lack identical counterparts in the serotype A or D strains studied so far. Indeed, two statistically robust and phylogenetically distinct haplotype groups were found only in serotype AD strains, not in the current serotype A or D sequences (Fig. 1). The first distinct cluster contains haplotype 4 (MAS92-0855-1 and MAS94-0351-1) and the second contains haplotypes 7, 8, 9 and

10 (see Table 2 and Fig. 1 for details). Sequencing additional serotype A or D strains at the *LAC* locus might reveal more identical sequences between serotype AD sequences and serotypes A or D. However, such results will only strengthen the rejection of hypothesis III.

Support for multiple hybrid origins of serotype AD strains (hypothesis IV)

The data are consistent with the hypothesis that serotype AD strains originated from multiple hybridization events. First, we observed two divergent clusters of sequences; one was highly similar to serotype A sequences and the other was similar to serotype D sequences (Fig. 1). Second, the two sequence clusters from serotype AD strains exhibited sequence diversities that were similar to those from serotype A or D (Table 3). Third, some sequences from serotype AD strains were identical to sequences from serotype A or D strains, which suggests that some hybridization events were very recent.

Based on the phylogenetic information from the *LAC* gene (Fig. 1; Table 2), there are at least three independent hybridization events among these 14 serotype AD strains. One event generated strains MAS92-0022 and MAS92-0793, with hybridization between strains containing haplotypes 1 and 6. A second hybridization event occurred between strains containing haplotypes 4 and 6; two strains are in this group, MAS92-0885 and MAS94-0351. A third hybridization event between strains containing haplotypes 1 and 7 gave rise to four strains: MAS93-0189, MAS92-0224, MAS93-0610 and MAS94-0244.

There are two possibilities for the origins of the other six serotype AD strains. The first possibility is that all six strains originated from independent hybridization events. The second possibility is that these six strains originated from the same three hybridization events mentioned above but have undergone mutation(s) and diverged. Both possibilities are compatible with the observed data, but the *LAC* sequence information is insufficient to provide statistical support for either.

Timing of hybridization events

Based on the *LAC* genealogical information, estimates for the time of hybridization are shown in Table 4. Major nodes in Table 4 are depicted in Fig. 1. The K2P distance was calculated between each serotype AD sequence and its closest counterpart from serotype A or D (Table 4). Estimates for time of hybridization from present ranged from 0 (present time) to 2.158 million years ago. All except three isolates had a low estimate of 0 based on the genealogical information (Table 4). Therefore, the results indicate that most hybridization events occurred relatively recently. Additional sequences from serotype A and D strains might reveal more sequences shared by strains of serotype AD and those from serotypes A or D. Such findings would reduce the

current upper estimates, thus providing stronger evidence for recent/current hybridization in *C. neoformans*.

DISCUSSION

Several studies have examined genetic variation within serotype AD strains and between serotype AD and other serotypes of *C. neoformans* (e.g. Brandt *et al.*, 1995, 1996; Xu *et al.*, 2000c; Boekhout *et al.*, 2001). These studies used MLEE, randomly amplified polymorphic DNA (RAPD), DNA sequence polymorphisms, DNA fingerprinting, and/or amplified fragment length polymorphisms (AFLP). Though these studies provided abundant evidence for the high variability of serotype AD strains, they have not been able to unambiguously distinguish between the possible hypotheses about the origin of serotype AD strains and their phylogenetic relationships with other serotypes. In this study, by applying a gene genealogical approach to analyse the highly polymorphic *LAC* locus, the four hypotheses were rigorously tested. The results indicate that multiple recent hybridization events between strains of serotypes A and D are responsible for the origin and current distribution of serotype AD strains.

Compared to the MLEE genotyping system, the *LAC*-gene-based sequence data are more informative for strain discrimination. Based on MLEE at 10 loci, the 14 strains analysed were classified into three MLEE genotypes (Table 1). In contrast, our analysis at the *LAC* locus revealed nine different diploid genotypes (Table 2). From the MLEE data, with the exception of one enzymic locus (6-phosphoglucanase dehydrogenase, 6PD), genotype information from the other nine loci showed that the three MLEE types (ET-5, 7 and 21) for serotype AD strains could be explained by a single hybridization event between ET-1 (serotype A) and ET-12 (serotype D) (see Brandt *et al.*, 1993, 1995). The differences among the three genotypes could be explained by differential loss of heterozygosity following hybridization. At the 6PD locus, there were three bands (alleles 2, 3 and 4) for the serotype AD strains: allele 2 was found only in serotype D strains and allele 4 only in serotype A strains. Because 6PD is a dimer, Brandt *et al.* (1993, 1995) proposed that allele 3 could be a combination of two monomers, one each from serotypes A and D. However, allele 3 from serotype AD strains has the same mobility as those from serotype B strains. Therefore, it is also possible that serotype B might be involved in the generation of serotype AD strains. Additional sequence information from this locus could resolve this question.

Our results demonstrated that each of the 14 strains had two distinct alleles at the *LAC* locus, but we cannot exclude the possibility that some strains might contain more than two alleles at this locus. However, for several reasons, such a possibility is unlikely to be common and should not affect our conclusion. First, for each of two strains, MAS93-0315 and MAS93-0610, two clones representing each of two *DpnI* restriction digest patterns

were sequenced. Each strain contained only two haplotypes. Second, assuming that a third putative class of haplotypes existed in comparable frequency (i.e. one-third) in each strain, among the total of 28 trials for the 14 strains, the chance of not picking and sequencing this putative class of haplotype is $[1 - (1/3)]^{28}$, or nearly impossible. Third, if additional putative alleles similar to those found in this study existed in some of the isolates, those alleles should not affect the interpretation of the data nor influence our conclusion that there were multiple hybridization events in the origin of serotype AD strains.

Aside from the MLEE data, other studies also suggested that loss of heterozygosity is likely to be common following hybridization between strains of serotypes A and D. First, Xu *et al.* (2000c) found homozygosity in all three nuclear genes in three serotype AD strains. In that study, differences in sequence were observed among these strains. For example, sequences of two serotype AD strains (CN110.97 and CN196.88) showed consistent clustering with serotype A strains, while the third (KW5) had different groupings, depending on the genes analysed (Xu *et al.*, 2000c). Second, Lengeler *et al.* (2001) showed that some serotype AD strains might be aneuploid and that homozygosity was found in three genes (CLA4, GPA1 and CNA1) in seven serotype AD strains examined. The results are consistent with the hypothesis that hybridization followed by differential loss of heterozygosity could be an important mechanism in generating genetic variation in *C. neoformans*. Interestingly, based on MLEE data, no loss of heterozygosity was observed among the serotype AD strains at the 6PD locus (Brandt *et al.*, 1995). Whether the maintenance of heterozygosity at this locus affords some fitness advantage for serotype AD strains is not known.

Our data are consistent with hypothesis IV and suggest at least three independent hybridization events are responsible for the 14 serotype AD strains analysed. Both bootstrap values (Fig. 1) and sequence diversity analysis (Table 3) provided strong statistical support for at least three hybridization events. By using serotype- and mating-type-specific primers, Lengeler *et al.* (2001) identified two potential hybridization events in a collection of seven serotype AD strains. However, for several reasons, the global distribution of serotype AD strains is likely to be the result of more hybridization events. First, all 14 strains analysed were isolated in the USA during a period of 2 years (1992–1994) (Brandt *et al.*, 1995, 1996). As such, these strains represent a spatially and temporally limited sample of the diverse, global populations of serotype AD strains. Second, geographic areas differ in the serotype distribution in *C. neoformans* (see Mitchell & Perfect, 1995; Table 1 in Xu *et al.*, 2000c). In geographic areas where both serotypes A and D are common, hybridizations between serotypes A and D are likely to happen more frequently. Unfortunately, very limited data are available for serotype distributions in the environment. Third, our study analysed only a fragment of one gene. Though this

locus (*LAC*) was the most polymorphic among the four loci we screened (Xu *et al.*, 2000c), analysis of other genes may reveal additional hybridization events.

This sequence analysis indicates that most hybridization events occurred relatively recently. Estimates for the times of hybridization ranged from 0 (present time) to 2.158 million years ago. All except three isolates had a low estimate of 0 (Table 4). The high extremes of the time ranges are probably overestimations. At present, the sequence database at the *LAC* locus for either serotype A or D strain is relatively small. With the expanding database of serotypes A and D, additional sequences identical or more similar to the serotype AD sequences could be uncovered and analysed. With such future studies, the upper limit for the time of hybridization could be reduced.

In conclusion, by sequencing and comparing a total of 28 fragments of the *LAC* gene from 14 serotype AD strains, we identified abundant sequence polymorphisms. Ten haplotypes and nine diploid genotypes were discovered among the 14 strains at this locus. All these alleles were from clinical strains capable of causing diseases, but any functional significance of these *LAC* haplotypes is unknown at present and awaits further investigation. Gene genealogical analysis rejects the hypotheses of either a single ancient or a single recent origin for serotype AD strains. Instead, our results are consistent with the hypothesis that multiple hybridization events between strains of serotypes A and D are responsible for the current distribution of serotype AD strains. Furthermore, most of these serotype AD strains originated from recent hybridization events. These results suggest a continuing dynamic evolution of the human pathogenic fungus *C. neoformans*.

ACKNOWLEDGEMENTS

Research support for J. Xu was provided by grants from McMaster University, the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Canadian Foundation for Innovation (CFI) and the Ontario Innovation Trust (OIT). G. Luo, T. Mitchell and R. Vilgalys were supported by Public Health Service grants AI 25783 and AI 44975 from the National Institutes of Health. All 14 serotype AD strains analysed in this study were collected through the CDC Cryptococcal Active Surveillance, a population-based active surveillance for cryptococcal disease. We thank Dr Tim Lott and Heather Yoell for their comments and to those who contributed to this surveillance effort.

REFERENCES

- Aulakh, H. S., Straus, S. E. & Kwon-Chung, K. J. (1981). Genetic relatedness of *Filobasidiella neoformans* (*Cryptococcus neoformans*) and *Filobasidiella bacillispora* (*Cryptococcus bacillispora*) as determined by DNA base composition and sequence homology studies. *Int J Syst Bacteriol* 31, 97–103.
- Boekhout, T., Theelen, B., Diaz, M., Fell, J. W., Hop, W. C. J., Abeln, E. C. A., Dromer, F. & Meyer, W. (2001). Hybrid genotypes in the pathogenic yeast *Cryptococcus neoformans*. *Microbiology* 147, 891–907.

- Brandt, M. E., Bragg, S. L. & Pinner, R. W. (1993). Multilocus enzyme typing of *Cryptococcus neoformans*. *J Clin Microbiol* **31**, 2819–2823.
- Brandt, M. E., Hutwagner, L. C., Kuykendall, R. J. & Pinner, R. W. (1995). Comparison of multilocus enzyme electrophoresis and random amplified polymorphic DNA analysis for molecular subtyping of *Cryptococcus neoformans*. *J Clin Microbiol* **33**, 1890–1895.
- Brandt, M. E., Hutwagner, L. C., Klug, L. A. & 9 other authors (1996). Molecular subtype distribution of *Cryptococcus neoformans* in four areas of the United States. *J Clin Microbiol* **34**, 912–917.
- Casadevall, A. & Perfect, J. R. (1998). *Cryptococcus neoformans*. Washington, DC: American Society for Microbiology.
- Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol* **17**, 368–376.
- Franzot, S. P., Salkin, I. F. & Casadevall, A. (1999). *Cryptococcus neoformans* var. *grubii*: separate varietal status for *Cryptococcus neoformans* serotype A isolates. *J Clin Microbiol* **37**, 838–840.
- Kabasawa, K., Itagaki, H., Ikeda, R., Shinoda, T., Kagaya, K. & Fukazawa, Y. (1991). Evaluation of a new method for identification of *Cryptococcus neoformans* which uses serologic tests aided by selected biological tests. *J Clin Microbiol* **29**, 2873–2876.
- Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**, 111–120.
- Kwon-Chung, K. J. (1976). A new species of *Filobasidiella*, the sexual state of *Cryptococcus neoformans* B and C serotypes. *Mycologia* **68**, 942–946.
- Kwon-Chung, K. J. & Bennett, J. E. (1984). Epidemiological differences between the two varieties of *Cryptococcus neoformans*. *Am J Epidemiol* **120**, 123–130.
- Kwon-Chung, K. J., Bennett, J. E. & Rhodes, J. C. (1982). Taxonomic studies on *Filobasidiella* species and their anamorphs. *Antonie Leeuwenhoek* **48**, 25–38.
- Kwon-Chung, K. J., Sorrell, T. C., Dromer, F., Fung, E. & Levitz, S. M. (2000). Cryptococcosis: clinical and biological aspects. *Med Mycol* **38** (Suppl. 1), 205–213.
- Lengeler, K. B., Cox, G. M. & Heitman, J. (2001). Serotype AD strains of *Cryptococcus neoformans* are diploid or aneuploid and are heterozygous at the mating-type locus. *Infect Immun* **69**, 115–122.
- Mitchell, T. G. & Perfect, J. R. (1995). Cryptococcosis in the era of AIDS – 100 years after the discovery of *Cryptococcus neoformans*. *Clin Microbiol Rev* **8**, 515–548.
- Ponce de Leon, G., Sattah, M., Graviss, E. A., Phelan, M., Brandt, M. E., Rimland, D., Hamill, R. & Hajjeh, R. A. (1999). Cryptococcosis surveillance: population-based trends, Atlanta and Houston, 1993–98. *Abstracts of the 37th IDSA meeting, Philadelphia, PA*, abstract no. 405.
- Swofford, D. L. (2001). *PAUP 4.0: Phylogenetic Analysis Using Parsimony*. Sunderland, MA: Sinauer Associates.
- Williamson, P. R. (1994). Biochemical and molecular characterization of the diphenol oxidase of *Cryptococcus neoformans*: identification as a laccase. *J Bacteriol* **176**, 656–664.
- Xu, J., Ali, R., Gregory, D., Amick, D., Lambert, S., Yoell, H., Vilgalys, R. & Mitchell, T. G. (2000a). Uniparental mitochondrial transmission in sexual crosses in *Cryptococcus neoformans*. *Curr Microbiol* **40**, 269–273.
- Xu, J., Ramos, A. R., Vilgalys, R. J. & Mitchell, T. G. (2000b). Clonal and spontaneous origins of fluconazole resistance in *Candida albicans*. *J Clin Microbiol* **38**, 1214–1220.
- Xu, J., Vilgalys, R. J. & Mitchell, T. G. (2000c). Multiple gene genealogies reveal recent dispersion and hybridization in the human pathogenic fungus *Cryptococcus neoformans*. *Mol Ecol* **9**, 1471–1481.

Received 12 June 2001; revised 13 September 2001; accepted 17 September 2001.