Evidence for a more recently evolved clade within a Candida albicans North American population

Timothy J. Lott and Mohammad M. Effat†

Author for correspondence: Timothy J. Lott. Tel: +1 404 639 2459. Fax: +1 404 639 3546. e-mail: tjl1@cdc.gov

Mycotic Diseases Branch, Division of Bacterial and Mycotic Diseases, National Center for Infectious Diseases, Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, GA 30333, USA

Candida albicans is diploid and displays a primarily clonal mode of reproduction. There is, however, evidence for meiosis and the degree to which this occurs in nature is unknown. Although random mating would act to obscure clonal lineages, previous studies have demonstrated that collections of North American isolates display three major partitions with no evidence of geographic clustering. To better understand the extent of sexuality and its role in the phylogeny of the species, a reference subset of 50 isolates representing this tripartite division was analysed using 1 minisatellite, 5 microsatellites (MSs) and 15 nuclear polymorphisms (NP). A total of 87 alleles were observed for 21 loci and 12/16 informative loci exhibited a departure from Hardy–Weinberg expectations ($G_2^a < 0.05$). We did not observe an absolute correlation between MSs and NP, although isolates with identical NP genotypes were correlated with a previously defined, predominant class (putative group I). The use of additional markers did not give increased support for the tripartite structure of the population. However, (9/19) group I isolates were found to be highly related, differing by only one or a few alleles. Designated subgroup A, the interpretation is that these isolates are related by descent and that they are of a more recent evolutionary origin, diverging from an ancestral group I clone. The reason for their relative abundance in the population is unknown; one possibility is that they may be under positive selection.

Keywords: genetics, population structure, SNPs, microsatellites

INTRODUCTION

The commensal yeast, Candida albicans, is a ubiquitous component of the normal human microflora. In non-immunocompromised individuals this species is often the agent of superficial epithelial infections. However, if normal immunity is reduced, C. albicans is capable of causing severe invasive disease. In the United States there has been an increase in immunosuppressed patient populations (Pfaller, 1995; Klein et al., 1984; Johnson et al., 1995). It is not surprising, therefore, that the number of C. albicans infections has been increasing and at present it is the eighth most common nosocomial bloodstream infection (Gerberding et al., 1999).

This increase in infection rate has resulted in a renewed interest in the genetics and epidemiology of this organism. C. albicans is diploid, with a haploid chromosomal number (N) of eight (Chu et al., 1993). An analysis of chromosomal stability has shown that the species undergoes relatively high rates of non-homologous mitotic recombination, translocation and aneuploidy (Thrash-Bingham & Gorman, 1992; Rustchenko-Bulgac, 1991). Although originally considered to reproduce as a clonal population (due to diploidy and the lack of a known sexual cycle), several initial analyses concluded that Hardy–Weinberg equilibrium could not be ruled out for some loci (Pujol et al., 1993; Graser et al., 1996). Additional work has yielded a consensus that C. albicans is primarily clonal, but with limited recombination (Forche et al., 1999). However the extent of sexuality in nature remains unclear. For example, a high degree of mitotic recombination/
Table 1. Summary statistics for C. albicans isolates

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
<th>R</th>
<th>S</th>
<th>T</th>
<th>U</th>
<th>V</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ne</td>
<td>48</td>
<td>31</td>
<td>13</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G^*</td>
<td>0.99</td>
<td>0.11</td>
<td>&lt;0.001</td>
<td>0.01</td>
<td>&lt;0.001</td>
<td>0.01</td>
<td>&lt;0.001</td>
<td>0.01</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>L^*</td>
<td>0.07</td>
<td>0.16</td>
<td>0.12</td>
<td>0.25</td>
<td>0.31</td>
<td>0.15</td>
<td>0.50</td>
<td>0.30</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
<td></td>
</tr>
</tbody>
</table>

*Calculated using 1000 simulated samples.

American population using alleles at loci ALS1 (Hoyer et al., 1995), CEF3 (Bretagne et al., 1997), ERK1 (Field et al., 1996). Anonymous loci were found by searching the C. albicans genome database (sequence data for C. albicans was obtained from the Stanford DNA Sequencing and Technology Center website at http://www-genome.stanford.edu/group/candida) for trinucleotide repeats of at least 6 (but less than approx. 25) repeat units in length.

METHODS

Isolates. From the 114 isolates previously described (Lott et al., 1999), 50 were randomly chosen for additional analysis.

Genetic analysis. MS ZNF1 was analysed using a previously described primer set (Field et al., 1996). Anonymous loci were found by searching the C. albicans genome database (sequence data for C. albicans was obtained from the Stanford DNA Sequencing and Technology Center website at http://www-genome.stanford.edu/group/candida) for trinucleotide repeats of at least 6 (but less than approx. 25) repeat units in length.

Primer sequences were based on non-repeated flanking regions giving predicted amplicons of between 100 and 300 bp. Forward primers were fluorescein-5’-end-labelled with either FAM or TAMRA. PCR products were purified in H2O, denatured and analysed on an automated sequencer (ABI310) using GENESCAN software (Applied Biosystems). Anonymous loci primers were chosen as follows. Locus A3, 5’-CAAGCTTATAGTGCTTACTA-3’ (F), 5’-CACAACCTTAGATACATCTCG-3’ (R); Locus A4, 5’-GTAATGATTACGGCAATGAC-3’ (F), 5’-AGAACGACGTGTACTATTGG-3’ (R); Locus A5, 5’-AGAACGACGTGTACTATTGG-3’ (R); Locus A5, 5’-TAGTCTTCTTATAGTAGTCA-3’ (F), 5’-CAGCAAGTCCAGGTGTCG-3’ (R). Locus A3 includes a nine-repeat unit of the trinucleotide TAA. Locus A4 includes 6 GAA repeats and locus A5 is a compound MS of 8 CAG repeats followed by a spacer and additional 3 and 4 tracts of CAG with a small intervening spacer. We observed that 8/50 strains exhibited four alleles at this locus and concluded that this region has been duplicated in these strains.

Primers and nomenclature for NP loci follow that of Xu et al. (1999b). Taq polymerase, dNTPs and buffers were from Boehringer Mannheim. Reaction conditions were as previously described (Lott et al., 1999). All PCR reactions were carried out in 50 µl reactions with 35 cycles (following an initial denaturation for 5 min at 95 °C) of 1 min at 95 °C, 1 min at 55 °C and 1 min at 72 °C, followed by 5 min at 72 °C for final elongation. For NP loci the PCR products were ethanol-precipitated, resuspended in 20 µl H2O and electrophoresed on 1.5 % agarose. Gels were stained with ethidium bromide and photographed.

aneuploidy would tend to reduce the number of heterozygotes in the population, thereby driving a departure from Hardy–Weinberg equilibrium. In addition the lack of phylogenetic constructs that are statistically supported (Tibayrenc, 1997; Vilgalys et al., 1997) is not expected for a primarily clonal organism. Finally, the recent discovery of a mating-type locus and its manipulation in vitro to allow cells of opposite ‘type’ to fuse, forming tetrapioids, would argue either for a cryptic sexual cycle or one that has been active in the evolutionary history of the species (Hull et al., 2000; Magee & Magee, 2000).

Despite the obscurity of clonality by meiosis, several studies of C. albicans population structure using phylogenetic approaches have demonstrated a tripartite division of non-epidemiologically related isolates. Pujol et al. (1997) showed parity between multilocus enzyme electrophoresis (MLEE), randomly amplified polymorphic DNA (RAPD) and Southern blot fingerprinting using a middle-repetitive probe. For these three groups, designated I–III, the PEPA allele was observed to be synapomorphic with group II (Pujol et al., 1997). This tripartite structure has been supported by other work, using different populations and methodologies (Schmid et al., 1999; McCullough et al., 1995), although it has not been universally observed (Xu et al., 1999a).

Recently, we have observed this division in a North American population using alleles at loci ALS1 (Hoyer et al., 1995), CEF3 (Bretagne et al., 1997), ERK1 (Field et al., 1996) and the nuclear insertion element IS1 (Mercure et al., 1993). Moreover, we established that the presence of the IS1 element is highly correlated with group III and, through inference, that this is the oldest of the three lineages (Lott et al., 1999).

In a recent study, Xu et al. (1999b) uncovered 15 anonymous single nuclear polymorphisms (NPs) which acted as independent codominant markers in their defined population. Interestingly, approximately 40% of the non-epidemiologically related strains were of a common genotype (Xu et al., 1999b). Consequently, we were interested in whether these additional loci would suggest that recombination is occurring within a subset of our previously described population of North American bloodstream isolates (Lott et al., 1999; Kao et al., 1999), and in determining the degree of correlation between microsatellites (MSs) and NPs. If found to be correlated, we were interested in whether these loci would give added support for clonality through a coalescence approach. In addition, we also report here on the finding of three new anonymous polymorphic MS markers, two of which were used in a population and phylogenetic analysis for a total of 21 loci.
Genetic analysis was performed using Arlequin 1.1 (Schneider et al., 1997) and Popgen 1.31 (Yang & Yeh, 1993). Phylogenetic analysis was performed using PHYLIP 3.5c (Felsenstein, 1992) and PAUP 4.0b (Swofford, 1998). For MS versus NP analysis, locus IS1 was grouped with the NPs. Similarity matrices were calculated as follows. For every given locus, if two isolates shared both alleles they were assigned a score of 10. If one allele was shared the value of 0–5 was assigned and for no common alleles a score of 0–0 was assigned. The summation was then divided by the total number of loci.

RESULTS

Population analysis

The summary statistics for the subset of 50 epidemiologically unrelated isolates are shown in Table 1. Loci ALS1, ERK1, CEF3 and IS1 have been described previously (Lott et al., 1999). The ALS1 gene encodes a 106 bp minisatellite; ERK1 and CEF3 are MSs, and IS1 is an optional insertion element within the 25S rRNA (Mercure et al., 1993). Locus zinc finger 1 is designated ZNF1 (Field et al., 1996). Anonymous loci 3 and 4 are trinucleotide repeats and, similar to the other MS alleles, are given as the number of base pairs of the amplified PCR product. For anonymous locus 5, in some strains there were four observed alleles per strain. We conclude that this gene has been duplicated in these strains and that the paralogue has conserved primer sites. There were 14 observed alleles at this locus, yet due to the duplication it was not included in the population and phylogenetic analysis. For loci F16N1.3–C2N8.1 the nomenclature for homo- and heterozygosity follows that of Xu et al. (1999b). The total number of observed alleles for each locus is given, followed by the effective allele number, denoted Ne (Schneider et al., 1997). This represents the number of equally frequent alleles that would be required to produce the same degree of homozygosity as observed in the population. We found that 5 of the 21 loci used in the population analysis were monomorphic, resulting in 16 informative loci. For the MSs we consistently observed that 1–3 alleles were predominant in the population, with lower frequencies for the remainder. The allelic sum was 87 and the mean Ne was 2.59 ± 2.0. The probability of Hardy–Weinberg equilibrium based on $\chi^2$ (Hartl & Clark, 1997) is given in Table 1. Fifteen of 16 (93%) loci were observed to have $P \leq 0.05$. During this analysis we observed that in some occurrences $|o - e| > e$, where $o$ is observed frequency and $e$ is expected frequency. We therefore calculated the likelihood ratio test ($G^2$) probability values (Yang & Yeh, 1993), which are given in Table 1. Under these criteria, 12/16 (75%) loci had $P \leq 0.05$. We conclude, therefore, that between 75 and 90% of the loci exhibited a significant departure from Hardy–Weinberg expectations. When all pairwise combinations of loci were tested for linkage disequilibrium (Schneider et al., 1997), none of the 120 $P$ values exceeded 0.05 (data not shown). This result suggests a clonal population structure.

We then calculated the probability of neutrality for each locus, using the Evans–Watterson test (Yang & Yeh, 1993). Under the assumption of neutrality, the upper and lower 95% confidence limits were calculated and the lower ($I_{nl}$) values are given in Table 1. In no case
was $P < 0.05$ (i.e. beneath the lower limit for neutrality), although for ALS1 the $L_{95}$ was 0.07. Therefore, we conclude that, with the possible exception of ALS1, each locus was neutral.

**Phylogenetic analysis**

A pairwise distance matrix was generated where each locus was given equal weight. From this, a phenogram was constructed using an unweighted pair group analysis (UPGMA) with arbitrary rooting and is shown in Fig. 1. The mean similarity equalled 69.5 ± 11.5%. Isolates previously identified as belonging to putative groups I, II and III (Lott et al., 1999) are shown to the right of the phenogram. Most group I isolates clustered as a unit and, within these, we observed a subset of more highly related isolates termed subgroup A. The complete genotypes of these nine strains are given in Table 2. Although no two strains in the 50 analysed had identical genotypes, these nine isolates differed by only one to a few alleles. As seen in Fig. 1, strains previously identified as belonging in groups I and III did tend to cluster, although there was some intermixing of the two groups. An analysis of the data using parsimony (Swoford, 1998) with heuristic searching methods was used to compare trees with those generated by UPGMA (data not shown). A bootstrap analysis failed to produce statistical evidence for a specific topology ($P > 70\%$). However, what was uniformly observed in the majority consensus trees was the grouping of nine highly related strains. This was also found in UPGMA analysis, as seen in Fig. 1.

To determine the extent of parity between MS and NP markers, the two datasets were analysed separately using both parsimony and distance-matrix methods. The same general conclusions were found as when the data were pooled; there were no trees observed with more than one or two minor branches with bootstrap values above 70%. The UPGMA phenograms for MSs and NPs (including locus IS1) are shown in Fig. 2. Absolute parity between the two marker sets was not observed; however, neither was there total discordance.

Significantly, subgroup A, that was observed to be related by MSs, was associated with two groups of NPs. As seen in Fig. 2 and Table 2, five isolates of subgroup A had a common NP genotype, while four isolates differed at loci P (F16N1.1) and W (C2N8.1), thus placing them in a separate NP clade.

**DISCUSSION**

**Clonality and recombination in C. albicans**

Our analysis of this collection of North American isolates lends additional support that clonality plays a major role in the genetics of *C. albicans* (Pujol et al., 1993; Forche et al., 1999). For 15 bi-allelic loci analysed, five were found to be of a single type. This was not unexpected in that Xu et al. (1999b) observed low frequency of polymorphisms at these loci, with the less frequent allele ranging from approximately 001 to 0.04%. Overall, the allelic frequencies for the remaining 10 NPs were in accordance with those previously described. While Xu et al. (1999b) observed 5/9 of these loci (not including IS1) to be in Hardy–Weinberg equilibrium, we observed 2/9 NPs based on either $\chi^2$ or $G^2$. The collection of isolates analysed in the present study represents a surveillance set of epidemiologically unrelated (i.e. one isolate per patient) strains from several different hospitals. Therefore, we did not elect to use clone correction in analysing allelic frequencies (Xu et al., 1999a). In addition, based on 23 loci (including A5 and its duplicated parologue), no two isolates had an identical genotype.

**Population structure: homoplasy and relatedness by descent**

In an analysis of 204 isolates using 15 NP markers, Xu et al. (1999b) observed that almost half (92/204) were of the same genotype. Using 14 of these NPs and IS1, we found that 24% (12/50) were of this same type, confirming this observation. Significantly, as shown in Fig. 2 (upper right-hand phenogram) this group is positively correlated with group I isolates (Fig. 2, upper-
Candida albicans population structure

Interestingly, as seen in Fig. 2 and Table 2, members of clade A can be divided into two groups depending on polymorphisms at two loci: F16N1.1 and C2N8.1. Both are Cfo1 polymorphisms and it is not known if, and how, they are related. It would logically follow, however, that these mutations were created following the formation of subgroup A. At present it is not known why members of subgroup A are relatively abundant (9/50 in the present population). One hypothesis is that they may be under positive selection. Since C. albicans is an obligate commensal, some form of host-mediated response would appear likely.

In this study we have observed that for at least some MSs, population structure can be correlated with NP mutations. This implies that they are not undergoing relatively higher rates of mutation; if so there would be no expected parity between marker sets. Rather, our interpretation is that neither MSs nor NPs alone may represent the true phylogenetic relationship between isolates and that this may be due to inherent limitations in each set. For MSs, it is known in S. cerevisiae and other species that large variations in mutation rates are observed depending on the type and length of repeat (Miret et al., 1998; Wierdl et al., 1997). For those presented in this study, the mutation rates are not known. Likewise for NPs, rate changes will depend on whether they are coding or non-coding, synonymous or non-synonymous, as well as other factors. In addition, both types of markers are influenced by physical linkage. For example, in the present study at least six of the NPs analysed are closely linked. Loci F16N1.1–1.6 are all located within a single 1800 bp fragment and only 12 genotypes are found in this region out of a possible 729 types. We believe that this would not be unexpected considering the physical proximity of the markers. Interestingly, there does not appear to be a positive correlation between physical distance and observed recombination in this region, as one might expect. Also of interest, in regard to the present findings, locus F16N1.1 is one of the two NP loci that subdivides group A. We believe, therefore, that additional approaches are needed to gain a better understanding of phylogenetic relatedness and to address such issues as the timing of clade divergence. Through this, we will better understand those factors influencing genetic diversity in this medically important species.

ACKNOWLEDGEMENTS

We would like to thank an anonymous reviewer for critical reading of the manuscript. This work was supported in part through a postdoctoral training fellowship awarded through the National Research Council of Egypt to M. M. E.

REFERENCES

between the presence of a self-splicing intron in the 25S rDNA of
Candida albicans

of polymorphic short and clustered coding-region microsatellites
to distinguish strains of Candida albicans. FEMS Immunol Med
Microbiol 15, 73–79.

Forche, A., Schonian, G., Graser, Y., Vilgalys, R. & Mitchell, T. G.
(1999). Genetic structure of typical and atypical populations of

National nosocomial infections surveillance (NNIS) system report,
Am J Infect Control 27, 520–532.

Graser, Y., Volosek, M., Arrington, J., Schonian, G., Presber, W.,
population structure of the human pathogen Candida albicans
exhibits both clonality and recombination. Proc Natl Acad Sci
USA 93, 12473–12477.


albicans ALS1: domains related to a Saccharomyces cerevisiae
sexual agglutinin separated by a repeating motif. Mol Microbiol
15, 39–54.

mating of the ‘asexual’ yeast Candida albicans in a mammalian host.

(1995). Emergence of azole drug resistance in Candida species
from HIV-infected patients receiving prolonged fluconazole

The epidemiology of candidemia in two United States cities:
results of a population-based active surveillance. Clin Infect Dis
29, 1164–1170.

Klein, R. S., Harris, C. A., Burkus, C., Small, B., Moll, B., Lesser, M.
the initial manifestation of the acquired immunodeficiency

Lott, T. J., Logan, D., Holloway, B., Fundygga, R. & Arnold, J.
(1999). Towards understanding the evolution of the human

Molecular epidemiology of the global and temporal diversity of

Candida albicans by construction of MTLα and MTLα strains.
Science 289, 310–313.

between the presence of a self-splicing intron in the 25S rDNA of
C. albicans and strains susceptibility to 5-fluorocytosine. Nucleic
Acids Res 21, 6020–6027.

dependent and sequence-specific expansions of CTG/CAG tri-
nucleotide repeats in Saccharomyces cerevisiae. Proc Natl Acad
Sci USA 95, 12438–12443.


(1997). Parity among the randomly amplified polymorphic
DNA method, multilocus enzyme electrophoresis, and Southern
blot hybridization with the moderately repetitive DNA probe Ca3


Schmid, J., Herd, S., Hunter, P. R. & 14 other authors (1999).
Evidence for a general-purpose genotype in Candida albicans,
highly prevalent in multiple geographical regions, patient
types and types of infection. Microbiology 145, 2405–2413.

Arlequin version 1.1: a software for population genetic data
analysis. Genetics and Biometry Laboratory, University of
Geneva, Switzerland.


contribute to chromosome length polymorphisms in Candida


Vilgalys, R., Graser, Y. & Mitchell, T. G. (1996). Response from

instability in yeasts: dependence on the length of the micro-
satellite. Genetics 146, 769–779.

Xu, J., Boyd, C. M., Livingston, E., Meyer, W., Madden, J. F. & Mitchell, T.
(1999a). Species and genotypic diversities and similarities of pathogenic

fragment length polymorphisms (RFLP) analyses reveal both
extensive clonality and local genetic differences in Candida