Towards understanding the evolution of the human commensal yeast *Candida albicans*

Timothy J. Lott, Brian P. Holloway, David A. Logan, Ruth Fundyga and Jonathan Arnold

Author for correspondence: Timothy J. Lott. Tel: + 1 404 639 2459. Fax: + 1 404 639 3546.

Allelic frequencies and relationships for one dimorphic locus and three unlinked polymorphic loci have been determined for 114 unrelated isolates of *Candida albicans*, including 14 laboratory reference strains and 50 strains from each of two geographic regions. Although there was no indication of geographical partitioning, there were significant correlations for specific allelic pairs among loci and little evidence that any alleles were in Hardy-Weinberg equilibrium. This gives additional support for the concept that the primary mode of genetic inheritance in this species is clonal, with other intracellular genetic events playing a lesser role in the creation of genomic diversity. Through inference of this and other known attributes of closely related *Candida* species, such as sequence analysis of IS1 and the ITS2 (internal transcribed spacer 2) region of the rDNA cistron, the deduced phylogeny suggests an evolutionarily recent origin for many frequently isolated strains. This finding will be of interest in the context of understanding pathogenicity and drug resistance in this human commensal yeast.

**Keywords:** *Candida albicans*, genetics, phylogeny, microsatellite, VNTR

**INTRODUCTION**

*Candida albicans* is a ubiquitous component of the normal human microflora and a frequent cause of superficial epithelial yeast infections. In immunocompromised individuals, however, this species is capable of invading host tissues, resulting in severe disseminated disease. *C. albicans* is a leading primary agent in nosocomial mycotic infections, the rate of which has been increasing during the past decade (Beck-Sague *et al.*, 1993). Further, there is evidence for emerging resistance to widely used antifungal (azole) compounds (Johnson *et al.*, 1995), and Southern blotting with cloned repetitive elements (Scherer & Stevens, 1988; Lockhart *et al.*, 1993). These have been used as typing schemes for a number of epidemiological investigations (Merz *et al.*, 1988; Diaz-Guerra *et al.*, 1997; Pfaller *et al.*, 1994).

To address the question of strain relatedness, approaches have focused on the analysis of allelic variation at multiple loci. This was first observed as phenotypic variation for growth with various carbon and nitrogen sources, tolerance to low pH, and other phenotypic characteristics (Odds *et al.*, 1983). More recent techniques include multilocus enzyme electrophoresis (MLEE; Caugant & Sandven, 1993), random amplified polymorphic DNA (RAPD; Lehmann *et al.*, 1992), ribosomal and mitochondrial polymorphisms.
these approaches, there appears to be fairly good agreement between them, and two to three subgroups within the species C. albicans have been identified (Clemons et al., 1997; Pujol et al., 1997).

With the advent of additional DNA sequence information, other loci have been identified that contain multiple alleles within the population. These include, but are not limited to, the ALS1 gene (a member of the agglutination-like substance gene family) (Hoyer et al., 1995), the ERK1 (protein kinase) locus (Metzgar et al., 1998) and CEF3 (translation elongation factor 3) (Breitagne et al., 1997). For the minisatellite ALS1, there is a 108 bp repeat unit encoding a 36 amino acid motif within the structural gene. An allele is therefore defined as the number of observed repeat units. For ERK1 and CEF3 the repeats are trinucleotide microsatellites; ERK1 encodes a polyglutamine tract within the structural gene, while CEF3 is a variable region within the putative promoter. Similar to ALS1, alleles are defined on the basis of repeat number (termed variable number of tandem repeats, or VNTRs). All three loci appear to be stable within a strain over many generations (Hoyer et al., 1995; Metzgar et al., 1998; Breitagne et al., 1997). In addition, recent evidence suggests that in the yeast Saccharomyces cerevisiae, at least certain classes of trinucleotide repeats, when present or below approximately 25 repeat units, have mutation rates that can approach the overall nuclear mutation rate (Miret et al., 1998). Both ERK1 and CEF3 have a mean number of repeat units below 25. Thus, due to the high number of alleles at each locus, these three loci may be informative not only for strain identification, but also in determining strain relatedness. In addition, Mercure et al. (1993) have determined that for a certain subgroup of C. albicans, strains harbour an IS1-like element within the 28S rRNA gene, contained on the largest chromosome (designated R). There is no indication that this element is mobile in C. albicans (Mercure et al., 1993). By definition this is not a VNTR, in that only a single copy of the element is known to be inserted. Although this insertion event could have occurred by multiple (and presumably rare) integrations, we believe that it is much more likely that these strains are related through a common ancestor, and that this may afford the possibility of finding true phylogenetic relationships between certain groups of strains. This has led us to determine allelic frequencies and relationships within a large population and to compare this with a smaller, previously characterized population (Pujol et al., 1997), with particular reference to those strains harbouring the IS1 element.

**METHODS**

**Yeast strains and maintenance.** Strains were grown and maintained on YPD (1%, w/v, yeast extract, 2%, w/v, peptone, 2%, w/v, glucose) medium at 35 °C as described previously (Lott et al., 1993). Routine manipulations followed the methods of Kaiser et al. (1994). One hundred C. albicans isolates were chosen at random from a surveillance study encompassing two metropolitan areas (San Francisco, CA, and Atlanta, GA). Each isolate was obtained from a blood sample representing an individual patient, and there was no epidemiological evidence of relatedness (Rees et al., 1998). A more complete description of all survey isolates will be presented elsewhere. Only three isolates in the present study were from HIV-infected individuals. Strains were initially speciated by Chromagar (Hardy Diagnostics) and API20C (BioMérieux Vitek) profiles. In addition, 14 isolates from a previous study (Pujol et al., 1997) were included, representing unrelated isolates from described subgroups.

**DNA isolation and PCR conditions.** Genomic DNA was extracted from 10 ml overnight cultures by using a Purgene Gram-positive and yeast DNA isolation kit (Gentra Systems) following the procedure of the manufacturer. Approximately 10 ng DNA was used per 50 μl PCR reaction. Taq polymerase, 10× buffer + MgCl₂ and PCR conditions followed the recommendation of the manufacturer (Boehringer Mannheim). Following an initial denaturation step at 95 °C, PCR conditions were 35 cycles of 1 min steps of 95, 55 and 72 °C, with a final elongation for 5 min at 72 °C. For ALS1 and CEF3, the annealing temperature was lowered to 50 °C.

**Primers and electrophoresis conditions.** IS1: 5'GGG AAT CTG ACT GTC TAA TTA A-3' (forward); 5'CTT GCC TGT GGT TTC GCT AGA T-3' (reverse). ALS1: 5'GGT GGT ACA AGT TCC ACT GCC-3' (forward); 5'CTG ACT GTC TAA TTA A-3' (reverse); 5'CCT CTT CCT TTC ATA TAG AA-3' (forward). For CEF3, the forward primer was end-labelled with TET (5'GGA TGA TCT CTT CCT TTC ATA TAG AA-3') and the reverse primer was labelled with FAM (5'GGT ACA ACT TCC ACT GCC-3'). The annealing temperature was lowered to 50 °C.

For CEF3, the forward primer was end-labelled with TET (5'tetrachlorofluorescin phosphoramidite; Glen Research) and for ERK1 the reverse primer was labelled with FAM (5'fluorescein; Glen Research). For IS1 and ALS1, following PCR amplification, the DNA was precipitated with 2.5 vols ethanol, dried and resuspended in 8 μl TE (10 mM Tris/HCl, 1 mM EDTA). The entire amount was then loaded into a single well. For IS1, gels consisted of 1.5% (w/v) agarose (Sigma), 1.0% (w/v) Nusieve GTG agarose (FMC). The PCR reactions were diluted fivefold in H₂O and analysed by an ABI system 310 (Applied Biosystems), using GENESCAN software (Applied Biosystems).

**Genetic analysis.** Parsimony analysis was conducted using PAUP version 4.0 (Swofford & Berlocher, 1987) running under GCG (Genetics Computer Group) with a UNIX systems alpha workstation (Digital Systems). Digenic disequilibria were calculated using Genpop (1.2) (Rousset & Raymond, 1995), and exact tests followed that of Weir (1990). An electronic copy of the data set can be obtained by contacting tjl1@cdc.gov.

**RESULTS**

To detect repeated elements and the presence/absence of IS1, oligonucleotide primers for PCR were designed to flank the variable regions and the resulting fragments sized and scored as an allelic type. In the case of ALS1, the allele number refers to the number of 108 bp repeats;
Table 1. Allelic frequencies and distributions

<table>
<thead>
<tr>
<th>Locus</th>
<th>Most common allele*</th>
<th>Next most common</th>
<th>Other alleles†</th>
<th>Total samples‡</th>
<th>More extreme§</th>
<th>Prob. of extreme¶</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALS1</td>
<td>16/83/0.43</td>
<td>13/13/0.07</td>
<td>98/0.50</td>
<td>5151</td>
<td>3999</td>
<td>0.00049</td>
</tr>
<tr>
<td>CEF3</td>
<td>135/61/0.31</td>
<td>126/57/0.29</td>
<td>78/0.40</td>
<td>17310</td>
<td>755</td>
<td>1.0×10^-14</td>
</tr>
<tr>
<td>ERK1</td>
<td>239/95/0.48</td>
<td>231/65/0.33</td>
<td>38/0.19</td>
<td>10290</td>
<td>5300</td>
<td>1.2×10^-8</td>
</tr>
<tr>
<td>IS1</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>15</td>
<td>10</td>
<td>0.02</td>
</tr>
</tbody>
</table>

* Shown as allele/allelic count/allelic frequency.
† Shown as allelic count/allelic frequency.
‡ No. of theoretically possible genotypic distributions.
§ No. of genotypic distributions less likely to occur than the observed data.
¶ Summed probability of all more extreme distributions.

these could easily be scored from the ‘ladder’ generated in each lane of the gel by using a lower-stringency PCR. In the case of CEF3 and ERK1, the alleles are indicated by the total size of the DNA fragments in bp. Finally, in the case of IS1 the presence or absence of the element was scored as ‘1’ or ‘0’, respectively. When two fragments were generated from a single strain, it was assumed to be heterozygous; likewise, a single fragment was scored as homozygous. Initially, 50 isolates each from two metropolitan areas were analysed separately. Surprisingly, for all three loci with variable size repeats (ALS1, CEF3 and ERK1), there was little evidence for a normal distribution around a mean size in the two populations. For example, in the case of CEF3, the highest frequency allele was 135 bp, the next most common were 126 and 130 bp, respectively. The remaining displayed a wide distribution in size range and were represented by one or a few isolates each. These frequencies were found to be similar in both populations and they did not differ significantly (Student’s t test, P > 0.01). Data from the two populations were pooled. The frequency for the presence of IS1 was also not significantly different between the two populations. The results for the combined analysis are given.

Agglutination-like substance 1 gene (ALS1, chromosome 6)

For the 108 bp repeat motif (Hoyer et al., 1995), 24 alleles were observed (n = 111). Copy numbers ranged from 4 to 37 and a common type (16/16) was observed at a frequency of 0.27. Sixty-two isolates had a unique type and the remaining types (19) were represented by 2–5 isolates each. The mean heterozygosity was 0.60.

Elongation factor 3 gene (CEF3, chromosome 5)

A total of 16 alleles (n = 112) were observed for the TTC/TTTC microsatellite located 308 bp upstream of the transcriptional start site of the CEF3 gene (Bretagne et al., 1997). The most frequent type (126/135) was found at a frequency of 0.419 and there were 19 types in total. Seven strains had unique types, and the frequencies of the other types were approximately the same. The mean heterozygosity was 0.77.

Protein kinase 1 gene (ERK1, chromosome 4)

For the clustered microsatellite encoding a Q/A polyamino acid tract (Field et al., 1996), 12 alleles were found (n = 113) giving rise to 16 allelic types. The most predominant allelic type (231/239) was present at a frequency of 0.40. The next most frequent type (239/239) was observed at a frequency of 0.16. Interestingly, type 231/231 was not observed. Eight isolates had unique types, and the heterozygosity was 0.57.

Insertion-like element 1 (IS1, chromosome R)

Flanking primers within the conserved 28S rDNA were used to amplify the 379 bp IS1 insertion element (Boucher et al., 1996), and three classes were observed: 0.196 (22 of 112 isolates) displayed a single fragment showing the presence of the element, 0.04 (5 of 112) were heterozygous, and the remaining majority did not contain the insertion. The mean heterozygosity was 0.04.

Composite analysis of loci

For all four loci collectively, 79 of 111 (0.71) strains where information was available displayed a unique type (i.e. each individual strain was unique) out of 85 total different types (6 types were represented by more than one strain). In addition, 22 of 111 (0.198) were monomorphic for all four loci and shared the allelic type: 16/16, 126/135, 231/239, 0/0 (ALS/ALS, CEF3/CEF3, ERK1/ERK1, IS1/IS1). For each locus separately, when the two most predominant alleles were considered, there were significant departures from the Hardy–Weinberg equilibrium under the assumption of random mating (χ² statistics, P<0.05 for randomness). It is interesting to note, however, that when all alleles were considered for each individual locus, three loci (CEF3, ERK1 and IS1) exhibited total frequencies of heterozygosity not significantly different from those expected...
Fig. 1. Parsimony analysis, where each represented strain is of a unique type. Clusters I, II and III correspond to those described by Pujol et al. (1997) and the isolate numbers and their corresponding positions are shown. Asterisks indicate isolates containing at least one copy of IS1. The arrow indicates the position of the 22 monomorphic isolates. The heuristically searched, majority consensus tree is unrooted.

from Hardy–Weinberg expectations. For the six possible pairwise combinations, digenic disequilibrium statistics were calculated, and all but one were observed to be highly significant (CEF3 to ALS1 was observed at $P = 0.039 \pm 0.023$). We then performed an exact test of Hardy–Weinberg expectations (Weir, 1990) for ALS1, CEF3 and ERK1, and a similar two-way exact test for IS1. For a given locus, a sample is defined by the allelic and genotypic counts. The same number of alleles (allelic counts) may be distributed in different ways such that it is theoretically possible to observe many different genotypic distributions for the observed number of alleles. A computer program was written to calculate the probability of observing each theoretically possible collection of genotypic counts, given the same allelic counts; it then sums the probabilities of all samples with the same allelic count that are less likely to be observed (more extreme) than the observed genotypic counts. Hardy–Weinberg expectations are based on the observed allelic counts and are calculated according to probability laws (Weir, 1990). Genotypic distributions in agreement with Hardy–Weinberg are expected. A probability near 1 indicates that the observed genotypic counts are consistent with random segregation, while a probability near 0 indicates a deviation from Hardy–Weinberg expectations. For the ALS1, CEF3 and ERK1 loci separately, alleles were grouped into three categories: most common, next most common, and all others (Table 1). None of the loci met Hardy–Weinberg expectations.

To determine allelic associations, parsimony analysis was performed for the 85 unique genotypes generated by eight alleles (per strain) and a majority consensus tree was compiled. All alleles were given equal weight, and the consensus tree is shown in Fig. 1. A bootstrap analysis to determine the statistical significance of the three identified clades failed to identify any nodes with confidence values above 70%. We therefore determined the genotypes of 14 unrelated strains from a previous study (Pujol et al., 1997), including four or five strains from each of the three clusters identified by using a different phylogenetic approach. These data were then combined with the original 85 unique strain types. We observed that the three clusters identified in the present study corresponded to the three clades previously identified, and that all isolates corresponded to their previous groupings (Pujol et al., 1997). Significantly, members of cluster III are associated (but not exclusively) with the presence of the IS1 element. These isolates also exhibit higher degrees of allelic variability at the other three loci. Conversely, cluster I includes the 22 isolates monomorphic for all loci (Fig. 1).

DISCUSSION

The use of allelic variation in defining population structure within a species or between closely related species has been well established, and the constraints to this type of approach have been recognized (McIntyre & Weir, 1997). It is possible to estimate, for example, the effective population size ($N_e$) needed when $x$ number of loci, each with $y$ number of alleles, are analysed (Selander & Levin, 1980). In the present work we have attempted to determine the number of alleles for three highly variable loci. We did not observe that any of the isolates were unstable (i.e. a particular strain-specific pattern would vary) supporting the observations of others (Bretagne et al., 1997; Field et al., 1996; Hoyer et al., 1995). In addition, we did not find more than two alleles per locus for any strain. This is particularly relevant in the case of ALS1, where it is known that this repeat is present in additional copies of a multigene
family (Hoyer et al., 1995). We did find, however, that for each locus there were two to three predominant alleles and a large number at low frequency. For our defined population \((n = 114)\), we observed a large number of unique alleles, and we infer from this that we have not uncovered the absolute number of alleles for the population at large. To do this (particularly in the case of ALS1) may require \(n\) to be in the thousands. However, our interpretation is that frequency versus size does not represent a normal distribution, and that although increasing \(n\) may determine the limits of size, the actual distribution pattern will not vary significantly.

Why then, are certain alleles predominant? One possibility is that these are selected for. In the case of ALS1, it has been suggested that the 36 amino acid repeat motif may have the function of determining relative 'adhesiveness' (Hoyer et al., 1995), and that certain domains of repeating units may have a role in optimal protein function. This may also be true for the translated microsatellite found in ERK1, although it is less obvious for the untranslated repeat found in the promoter of CEF3 (Bretagne et al., 1997). Our interpretation, however, based on what is known concerning microsatellites in other eukaryotic systems (Field & Wills, 1996), is that within certain limits repeat size is neutral, and that size variation is generated at a certain rate through specific genetic mechanisms (Strand et al., 1994). This has led us to consider whether certain allelic combinations are random (suggesting something like meiosis) or more tightly associated, as would be the case in a primarily clonal mode of inheritance.

In the present study we have observed little evidence for allelic equilibrium, either within a given locus or between loci. We conclude, therefore, like others (Graser et al., 1996; Pujol et al., 1993) that the predominant form of inheritance in C. albicans is clonal, and we hypothesize that departures from strict clonality may be due to other known mechanisms that create genetic diversity in this species (Thrash-Bingham & Gorman, 1992).

At present there is no evidence that VNTRs are conserved to any degree at the interspecies level within the genus Candida. However, due to the previously demonstrated stability of these alleles over many generations and the observations from this study, relatedness owing to allelic size may be useful in developing a phylogeny for strains within the species C. albicans. For multiallelic loci, phylogenetic analysis using distance matrix algorithms is complex and involves many assumptions concerning the evolution of a particular allelic state. We therefore chose a parsimony approach and made no assumptions (such as 'weighted states'). The problem, however, is that when more than a few alleles are present, the number of possible topological trees becomes very large and the statistical evidence for a particular grouping (the 'bootstrap') becomes increasingly small. Nevertheless, when a subset of strains from a previous analysis (Pujol et al., 1997) were included, all of the isolates clustered in the same manner as previously reported. It should be noted that in the former study an entirely different set of polymorphisms were analysed by using a different phylogenetic algorithm (Pujol et al., 1997). We believe, therefore, that the fundamental topology of the three clusters is correct. Although the majority of consensus trees showed the order of groups as III, I, II, as shown in Fig. 1, there were a significant number of trees generated with a III, II, I order, and the precise assignment must therefore be considered as tentative.

There are two other aspects of Fig. 1 that we believe are important for deciphering the true phylogeny of C. albicans. The first is that without exception strains harbouring the IS1 element belong to group III (although the converse, that all group III strains contain the IS1 element, is not necessarily true). Second is the observation that groups I and II show, in general, less variability in allelic type than group III isolates. For example, in the present study 22 isolates in group I were monomorphic for all loci studied. If it is assumed that allelic diversity is a function of mutational rate (Rousset & Raymond, 1995), our interpretation, shown in the cladogram depicted in Fig. 2, is that group III isolates represent an evolutionarily older and more diverse class, while groups I and II are of more recent origin and show less variability. In addition, the presence of IS1 in group III isolates is also consistent with what is known about
this mobile element in other closely related Candida species. For example, in C. albicans, the element has been truncated from a more typical full-length sequence by the deletion of two stem-loop regions (Mercure et al., 1993). For Candida stellatoidea, the element is the same length as in C. albicans and over 99% identical in sequence (Boucher et al., 1996). In Candida dublinensis, the element is full-length and in the conserved regions shares a high degree of homology with C. albicans, although lower than the albicans–stellatoidea homology (Boucher et al., 1996). Finally, for the related species Candida parapsilosis, an analysis of 42 independent isolates has not demonstrated the presence of the element (T. J. Lott, unpublished). Our interpretation of these results is also consistent with 18S and ITS2 sequence homology data (Barnes et al., 1991; Santos et al., 1997), as well as other genetic aspects of these related species. For example, it has been suggested that from a common ancestor sharing a CUG' codon change (Pesole et al., 1993), species such as C. parapsilosis and Candida tropicalis diverged (Fig. 2, position 'A'). Some later time there was the gain of IS1 in C. albicans, followed by the divergence of C. dublinensis (Fig. 2, position 'B'). At some later point following the truncation of IS1, C. stellatoidea diverged (Fig. 2, position 'C'). Most recently, in C. albicans there has been the divergence of groups I and II and the loss of IS1 (Fig. 2, position 'D'). An interesting correlate to this hypothesis is that C. dublinensis and C. albicans have undergone concerted evolution, in that phenotypically they cannot be distinguished (Boucher et al., 1996).

Although we cannot exclude the possibility of population sample bias, the C. albicans strains used in this study have no known epidemiological relationship, and there is no evidence for geographical partitioning by strain type. Thus, it will be of interest to determine why a large number of strains represented in this population are of a common lineage (i.e. monomorphic for all loci studied), and whether these common alleles have a selective advantage in the host. However, other well-established factors may also be present in defining population structure (McIntyre & Weir, 1997). The establishment of clonal lineages will also aid in our understanding of emerging phenotypes, such as drug resistance, where there is often evidence for pleiotropic mutations of unknown genetic origin (Sanglard et al., 1995). With the advent of additional genomic sequences, it is hoped that other markers will be uncovered that evenly cover all eight chromosomes. This will clearly facilitate a genetic understanding of this medically important yeast and perhaps help address other questions, such as the time framework for the evolution of distinct subgroups.

ACKNOWLEDGEMENTS

We thank the Centers for Disease Control and Prevention’s antifungal surveillance group and Dr David Soll for providing C. albicans strains. We would also like to thank Cheryl M. Elie for technical support and Travis Benton for graphic assistance. This work was supported in part by an American Society for Microbiology faculty fellowship awarded to D.A.L.; R.F. was supported through NIH training grant AI07373.

REFERENCES


Evolution of C. albicans


