Functional specificity of *Candida albicans* Als3p proteins and clade specificity of *ALS3* alleles discriminated by the number of copies of the tandem repeat sequence in the central domain

Soon-Hwan Oh, Georgina Cheng, Jennifer A. Nuessen, Robert Jajko, Kathleen M. Yeater, Xiaomin Zhao, Claude Pujol, David R. Soll and Lois L. Hoyer

*Department of Veterinary Pathobiology, University of Illinois at Urbana-Champaign, Urbana, IL 61802, USA
Department of Biological Sciences, The University of Iowa, Iowa City, IA 52242, USA*

*Candida albicans* strain SC5314 contains two *ALS3* alleles, which differ in sequence with respect to the number of copies of the 108 bp tandem repeat sequence within the central domain of the coding region. One allele (*ALS3(12)*) has 12 tandem repeat copies while the other (*ALS3(9)*) has 9 copies. Wild-type *C. albicans* (*ALS3(12)/ALS3(9)*) and those containing various *ALS3* alleles (*ALS3(12)/als3Δ(9), als3Δ(12)/ALS3(9) and als3Δ(12)/als3Δ(9)*) were assayed for adhesion to monolayers of cultured vascular endothelial and pharyngeal epithelial cells. These assays showed obvious adhesive function for the larger Als3p protein, compared to a minor contribution to adhesion from the smaller protein. These functional differences in strain SC5314 prompted examination of *ALS3* allelic diversity across the five major genetic clades of *C. albicans*. This analysis focused on the number of copies of the tandem repeat sequence within the central domain of the coding region and showed a range of alleles encoding from 6 to 19 tandem repeat copies. Clades differed with respect to prevalent *ALS3* alleles and allele distribution, but were similar for the mean number of tandem repeat copies per *ALS3* allele. Analysis of allelic pairing showed clade differences and the tendency for *C. albicans* strains to encode one longer and one shorter *ALS3* allele. The allelic variability observed for *ALS3* and its functional consequences observed in strain SC5314 highlight the importance of understanding *ALS* allelic diversity in order to draw accurate conclusions about Als protein function.

INTRODUCTION

The *ALS* (agglutinin-like sequence) gene family of *Candida albicans* encodes eight large cell-surface glycoproteins, some of which function in adhesion to host cells (Gaur & Klotz, 1997; Hoyer, 2001; Fu et al., 2002; Zhao et al., 2004). The eight *ALS* genes share a common three-domain structure (Hoyer, 2001). The 5’ domain is 1299 to 1308 bp in the various *ALS* genes and is 55 to 90% identical across the family. The central domain of each *ALS* gene is composed entirely of tandemly repeated copies of a 108 bp motif. The *ALS 3’* domain is relatively variable in length and sequence, but in each gene, encodes a Ser/Thr-rich sequence that, in the mature Als protein, is heavily glycosylated (Kapteyn et al., 2000; Hoyer, 2001). Although allelic sequence variability exists in each region of the *ALS* genes (Hoyer, 2001; Zhang et al., 2003; Zhao et al., 2003), when considering the entire *ALS* family, the central tandem repeat domain contributes the greatest differences between *ALS* alleles as a result of variability in the number of copies of the 108 bp sequence.

Recently, adhesive function was demonstrated for Als3p, a protein that is produced on *C. albicans* germ tubes and hyphae (Hoyer et al., 1998; Zhao et al., 2004). Disruption of *ALS3* results in cells with reduced adherence to vascular endothelial and buccal epithelial cells, as well as a marked reduction in destruction of epithelial cells in the reconstituted human epithelium *in vitro* model of candidiasis (Zhao et al., 2004). Allele-specific lengths of the tandem repeat domain in *ALS3* were reported previously (Hoyer et al., 1998; Zhao et al., 2004). In strain SC5314, which is used most often for constructing *C. albicans* mutants (Fonzi & Irwin, 1993), DNA sequencing verified that one *ALS3* allele encodes 9 copies of the tandem repeat sequence, while the other encodes 12 copies (Zhao et al., 2004). Similar
variability within the tandem repeat domain has been reported for ALS1 (Lott et al., 1999), where the number of repeat sequences per allele ranged from 4 to 37, with 16 copies the most frequent. Analysis of the tandem repeat domain of ALS7 detected alleles that varied from 1 to 33 tandem copies, with 17 copies the most frequent (Zhang et al., 2003). Both of these studies demonstrated that some ALS alleles are more common than others and that, among alleles of a specific ALS gene, variability in the number of tandem repeat copies can be large.

During efforts to functionally characterize Als3p, we created C. albicans strains heterozygous for each ALS3 allele. In the studies presented here, we demonstrate variability in adhesion of these strains, with greater adhesion associated with the presence of the larger ALS3 allele. These results suggested functional differences between Als3p proteins in strain SC5314 depending on the number of tandem repeat copies in the central domain. This observation prompted examination of the distribution of ALS3 alleles, discriminated by the number of tandem repeat copies, among a larger population of C. albicans strains. For this work, we focused on genetically distinct groups of C. albicans that were assigned to five major clades (I, II, III, SA and E) by DNA fingerprinting with the species-specific complex probe Ca3 (Pujol et al., 1997, 2002; Soll, 2000; Blignaut et al., 2002; Soll & Pujol, 2003). One of these clades, SA, is relatively specific to South Africa, while a second clade, E, is relatively specific to Europe. A third clade, II, appears to be absent in the Southwest USA. Recently, it was demonstrated that all tested strains of C. albicans naturally resistant to 5-fluorocytosine (MIC > 32 μg ml⁻¹) and the great majority of strains less susceptible to 5-fluorocytosine (MIC ≥ 5 μg ml⁻¹) are members of clade I (Pujol et al., 2004) and that clade-specific 5-fluorocytosine resistance is due to a single nucleotide change in the FUR1 gene (Doddson et al., 2004). The maintenance of deep-rooted clades side by side in the same geographical area, as well as the relative uniformity of the less susceptible 5-fluorocytosine phenotype in clade I, has led to the suggestion that recombination may occur frequently within a clade, but infrequently between clades (Soll & Pujol, 2003; Pujol et al., 2004). This hypothesis suggested that other clade-specific differences, perhaps for ALS3 alleles, are likely to exist across the general population of C. albicans strains.

Using PCR-based methods to discriminate the number of tandem repeat copies in the central domain of the ALS3 alleles of representative strains in each of the five major C. albicans clades, we demonstrate dramatic clade-specific differences in the distribution of ALS3 alleles. Our results also reveal that, on average, the two ALS3 alleles of a strain encode one ‘long’ and one ‘short’ Als3p, suggesting the possibility of a functional role for the long and short combination. These data, combined with observed differences in adhesion contributed by shorter and longer ALS3 alleles in strain SC5314, suggest that the variability in ALS3 alleles among the five clades identified in this study may also reflect functional differences.

METHODS

Candida strains. The collection of 196 isolates used in the present study was obtained from three previously analysed populations (Pujol et al., 1997, 2002; Blignaut et al., 2002) and included 84 isolates from the United States and Canada, 71 from South Africa, 25 from Europe, 8 from South America, and 4 from Turkey and Israel. The collection included 51 clade I isolates, 47 clade II isolates, 40 clade III isolates, 36 clade SA isolates and 22 clade E isolates.

Methods for constructing strains heterozygous for each ALS3 allele and als3Δ/als3Δ null strains were described by Zhao et al. (2004). The set of strains was constructed twice in its entirety, starting with strain CAI4 (Fonzi & Irvin, 1993). Strain CAI12, used as a control for the studies described here, was created by reintegration of URA3 into its native locus in the CAI4 background (Porta et al., 1999). Both CAI4 and CAI12 were generous gifts from Dr William Fonzi (Georgetown University, Washington, DC, USA). C. albicans SC5314 encodes a larger ALS3 allele (ALS3(12)); 12 copies of the 108 bp tandem repeat sequence in the central domain; GenBank accession no. AY223552) and a smaller ALS3 allele (ALS3(9); 9 tandem repeat copies; GenBank accession no. AY223551). Strains resulting from two independent construction efforts were 1704 and 1893 (ALS3(12)/als3Δ(9) URA3), 1702 and 1897 (als3Δ(12)/als3Δ(9) URA3), and 1843 and 1954 (als3Δ(12)/als3Δ(9) URA3). There was no statistically significant difference between any of the strains for growth rate, the ability to form germ tubes in RPMI 1640 medium, or cellular aggregation. Methods described by Zhao et al. (2004) were used to evaluate these characteristics.

C. albicans adhesion to human cell monolayers. The method to measure adherence of C. albicans strains to human umbilical vein endothelial cells was described by Zhao et al. (2004). These assays were conducted in a six-well tissue culture plate format using modifications of the method described by Ibrahim et al. (1995). Six-well plate assays were also conducted to measure adhesion of C. albicans to human pharyngeal epithelial cells. The epithelial cell line was purchased from the American Type Culture Collection (HTB-43, also called FA-DU) and grown in Minimal Essential Medium (MEM; Gibco) containing 10% fetal bovine serum (Gibco). Growth medium was supplemented with 2 mM l-glutamine and 1 mM sodium pyruvate; no antimicrobials were added. Confluent epithelial cell monolayers were grown in six-well tissue culture plates as described for the endothelial cell assay (Zhao et al., 2004). C. albicans strains were cultured in RPMI 1640 to form germ tubes prior to the adhesion assay as previously described (Zhao et al., 2004) and 500 cells with germ tubes added to each assay well. Following incubation, the monolayer was washed once with 5 ml Dulbecco’s phosphate-buffered saline (no calcium or magnesium). Agar overlay of each well, viable counts of C. albicans c.f.u. and calculation of percentage adherence were carried out as described before (Zhao et al., 2004). In a typical assay, wild-type CAI12 showed approximately 150 to 200 adherent c.f.u. per well. Six assay-wells were used for each strain in three separate experiments. The statistical significance of results was evaluated using a mixed-model analysis of variance (PROC MIXED in SAS). Separation of means was performed using the LSMEANS option. Specific comparisons between strains or combinations of strains were calculated using the ESTIMATE function.

ALS3 genotyping. Genomic DNA was extracted from the C. albicans strains using a previously described method (Hoyer et al., 1995). The size of the tandem repeat domain in each ALS allele was determined by PCR using two independent primer pairs. Use of two primer pairs provided an additional control for the accuracy of the
results. Each primer pair contained one that annealed 5' and another 3' of the tandem repeat domain. The first primer pair was ALS3GenoF (5'-ACC TTA CCA TTC GAT CCT AAC C-3') and ALS3GenoR (5'-GAT GGG GAT TGT GAA GTG G-3'). The second primer pair was ALS3GenoF2 (5'-CCA CAA CAC ATA CTA ATC CAA CTG A-3') and ALS3GenoR2 (5'-TGT AGA CCA CAA AGT TGT ATG GTT G-3'). Taq polymerase (Invitrogen) was used with both primer pairs. Reactions with the first primer pair (ALS3GenoF and ALS3GenoR) used Invitrogen Taq polymerase buffer with 1 mM MgCl2. Reactions were heated for 5 min at 94°C followed by 35 cycles of 94°C (30 s), 57°C (30 s) 72°C (3 min). A final 72°C (7 min) extension completed the reaction. The second primer pair (ALS3GenoF2 and ALS3GenoR2) was used under similar conditions except for a difference in buffer (10 mM Tris/HCl, pH 8.8, 25 mM KCl, 1.5 mM MgCl2) and annealing temperature (65°C). PCR products were separated on 3-5% polyacrylamide-Tris/borate/EDTA (TBE) gels and stained with ethidium bromide. Amplicon sizes were determined by comparison with the 1 kb ladder (Invitrogen) and with ALS3 amplification products from strain SC5314 genomic DNA. In this strain, allele size was determined by DNA sequencing. Alleles have either 9 (GenBank accession no. AY223551) or 12 (AY223552) copies of the 108 bp tandem repeat in the central domain. The two pairs of ALS3 genotyping PCR primers were verified against Southern blots of genomic DNA from a variety of C. albicans strains (J. A. Nuesen & L. L. Hoyer, unpublished data). Genomic DNA was digested with either HindIII/NcoI or FokI/ApaI and hybridized with an ALS3-specific fragment from the 3' end of the gene that was contained within the restriction sites. The allelic tandem-repeat copy number and allelic combination derived for each strain was the same when determined using Southern blots or the PCR primer pairs described above.

**RT-PCR reactions.** RT-PCR analysis used the method described by Green et al. (2004). C. albicans cells were grown overnight in YPD medium (per litre: 10 g yeast extract, 10 g peptone, 20 g glucose), washed in PBS, counted with a haemocytometer and released into prewarmed RPMI 1640 medium at a density of 5 x 106 cells ml⁻¹. These cultures were grown for 1 h at 37°C and 200 r.p.m. shaking; at this point, germ tubes were visible microscopically. Cells were harvested by filtration, flash-frozen in dry ice/ethanol and stored at −80°C. RNA was extracted using a hot phenol method (Collart & Oliviero, 1993), DNase treated (Ambion) and quantified spectrophotometrically. Synthesis of cDNA used 2 μg total RNA and the SuperScript First-Strand Synthesis System for RT-PCR (Invitrogen), primed with random hexamers. One-tenth volume was added to PCR reactions using the ALS3GenoF and ALS3GenoR primers and amplified as described above. Equal quantities of RNA without addition of reverse transcriptase was run to ensure that the observed products were not amplified from residual genomic DNA.

**Statistical methods.** The distributions of the number of tandem repeat copies in the different clades and the distributions of the difference in the number of tandem repeat copies per strain in the different clades were evaluated using the non-parametric Kolmogorov–Smirnov test. This approach was selected because the data were not distributed normally. The non-randomness of allelic combinations was calculated using Fisher’s Exact test. Allele frequencies were used to calculate the expected frequencies of the diploid genotypes under the null hypothesis that combinations of alleles were random. The observed and expected numbers were then compared using Fisher’s Exact test. Fst statistics were assessed using the MULTILocus 1.3 software package, available at http://www.agapow.net/software/multilocus/ (Agapow & Burt, 2001). The program calculates Weir’s formulation of Wright’s Fst (Weir, 1996). The calculated statistic, 6, is an approximation of the Fst. Fst statistics are used in population genetics studies as an estimate of the level of subdivision among populations to test their differentiation. An Fst of 0 means that there is no subdivision (the populations do not show any statistical difference), whereas an Fst of 1 means that populations are completely isolated genetically (the populations have nothing in common). In our study, Fst statistics were used to test if ALS3 gene frequencies reflected the genetic divergence shown by Ca3 fingerprinting among C. albicans clades and to determine whether clade differences in the study were a function of the geographic origin of the strains analysed.

**RESULTS**

**Variability in adhesion function of proteins encoded by the different ALS3 alleles in strain SC5314**

Phenotypic analysis of two independently constructed als3Δ/als3A mutant strains showed greater than 50% reduction in adherence to vascular endothelial cell monolayers compared to the wild-type strain CAI12 and a 60% reduction in adherence to buccal epithelial cells in suspension (Zhao et al., 2004). To assess the effect of the different ALS3 alleles on adhesion, heterozygous strains expressing only the smaller ALS3 allele of strain SC5314 (ALS3(9); 9 tandem repeat copies) or only the larger ALS3 allele of strain SC5314 (ALS3(12); 12 copies) were created. The presence of only one functional ALS3 allele in each strain was demonstrated by RT-PCR (Fig. 1). The als3Δ/ALS3 strains were compared to wild-type (CAI12; ALS3(12)/ALS3(9)) and als3Δ/als3A strains in adhesion assays using monolayers of human umbilical vein endothelial cells (HUVEC; Fig. 2a) or pharyngeal epithelial cells (FaDu; Fig. 2b). For both cell types, strains expressing only the larger ALS3 allele (ALS3(12)/als3A(9)) showed the same adherence as the wild-type control (P=0.76 for HUVEC; P=0.95 for FaDu). When only the smaller allele was present, the als3Δ(12)/ALS3(9) strains were significantly less adherent than the wild-type control (P=0.0015 for HUVEC; P=0.030 for FaDu) and the als3A(9)/ALS3(12) strains (P=0.0001 for HUVEC; P=0.0084 for FaDu). Adhesion of the als3Δ(12)/ALS3(9) strains was significantly different from the als3Δ(12)/als3A(9) strains (P=0.0032 for HUVEC; P<0.0001 for FaDu), suggesting that there was a contribution to adhesion from the smaller Als3 protein. For both HUVEC and FaDu cell monolayers, the als3Δ(12)/als3A(9) strains were significantly different in adhesion from wild-type and ALS3(12)/als3A(9) (P<0.0001 for all comparisons). These results demonstrate the greater contribution to adhesion by the larger Als3 protein for both endothelial and epithelial surfaces.

Comparison of the nucleotide sequences of the larger ALS3 allele (GenBank accession no. AY223552) and smaller ALS3 allele (GenBank accession no. AY223551) revealed that, outside of the tandem repeat domain, there were six nucleotide changes between the two sequences. Two of these were in the 5’ region, but did not result in alteration of the amino acid sequence of the Als3p N-terminal domain. Four nucleotide changes were in the ALS3 3’ region. One did not result in an amino acid change in
the C-terminal domain, and one caused a conservative replacement (Asp/Glu). The remaining sequence differences resulted in a Phe/Val and a Gln/Pro substitution. The C-terminal domain of Als proteins is so heavily glycosylated (Kapteyn et al., 2000) that these amino acid replacements are highly unlikely to alter protein structure. The near-perfect allelic sequence conservation in N- and C-terminal domains suggested that the observed differences in adhesion between the Als3p proteins from strain SC5314 were due to the length of the central tandem repeat domain. These studies demonstrated the importance of ALS3 allelic variability in assessing protein function and showed how considerably different conclusions can be reached depending on which allele was studied. The difference in the adhesive contributions of the two Als3p proteins in strain SC5314 prompted work to define ALS3 alleles based on the length of the tandem repeat domain across a larger population of C. albicans strains.

Range of alleles based on number of tandem repeat copies

A collection of strains was randomly selected to represent the five major C. albicans clades (Soll & Pujol, 2003). This collection included 51 strains from clade I, 47 strains from clade II, 40 from clade III, 22 from clade E and 36 from clade SA. The proportions of the different ALS3 alleles in the general C. albicans population were assessed by PCR using primers that annealed 5’ and 3’ of the tandem repeat domain to determine the number of tandem repeat copies in each allele. The number of copies of the tandem repeat sequence was calculated for each of the two alleles in the 196 strains, resulting in 392 analysed ALS3 alleles. The range of tandem repeat copy number was between 6 and 19 (Fig. 3, Table 1). The alleles did not fall into a Poisson distribution. The majority of the alleles (78%) had either 10 copies (28%), 11 copies (22%) or 13 copies (28%) of the repeat sequence. The remaining 22% of the alleles represented all other categories (Fig. 3, Table 1). Alleles that contained 12 or 14 copies of the tandem repeat sequence were under-represented compared to the other categories (Fig. 3). Overall, only 10% of the total alleles had 9 tandem repeat copies or fewer, while 82% of the alleles had between 10 and 13 repeat copies. Larger repeat copy numbers (>14) were less common, representing only 8% of the total alleles.

Clades differ in prevalent ALS3 alleles

ALS3 genotyping data revealed distinct differences between clades (Table 1). The great majority of ALS3 alleles in clade

---

**Fig. 1.** Ethidium-bromide-stained RT-PCR products to demonstrate ALS3 allele expression in various C. albicans strains. C. albicans strains with two, one or no ALS3 alleles were grown in RPMI 1640 medium for 1 h to induce expression from the ALS3 promoter. RNA from each of the strains noted above the image was reverse transcribed and amplified by PCR using primers ALS3GenoF and ALS3GenoR. Products were separated on a 3·5% polyacrylamide/TBE gel, stained and photographed. Control reactions to which reverse transcriptase was not added were run as controls to ensure that amplified products were not derived from genomic DNA. The position of molecular size markers (in kb) is noted on the left of the image and the identity of the ALS3 alleles is noted on the right.

**Fig. 2.** Adhesion assay data for C. albicans strains on human umbilical vein endothelial cells (a) and pharyngeal epithelial cells (b). Adhesion assay data were normalized with adhesion of the wild-type control (CAI12) designated 100%. Means and standard errors for each of the tested strains are depicted in the histogram.
I isolates contained either 10 (44%) or 13 (47%) tandem repeat copies (Table 1). No other allelic category in clade I contained more than 2% of clade I alleles. Clade I was the only group that had alleles with more than 15 repeat copies, although such alleles were observed rarely. The bias toward alleles containing 10 or 13 tandem repeat copies is evident in the dendrogram model in Fig. 4, in which each allele is colour-coded and represented according to its proportion in each clade. It is also evident in the dendrogram model that clade E was composed mainly of ALS3 alleles with either 11 or 13 tandem repeat copies. Compared to clades I and E, the other clades included a greater variety of alleles (Table 1, Fig. 4).

Comparing the proportions of ALS3 alleles by clade demonstrated that common alleles in one clade were under-represented in others. For example, alleles with 10 tandem repeat copies were common in all clades except E, where none were detected. ALS3 alleles with 11 tandem repeat copies were nearly absent in clades I and SA, while they were among the most prevalent in the other clades. Clade SA was the only one that did not have a significant proportion of alleles with 13 tandem repeat copies. Clade SA was

![Fig. 3. Frequency distribution of the number of tandem repeat copies present in the 392 ALS3 alleles tested in this work.](image)

![Fig. 4. Dendrogram model to show the relationship between clades and the frequency of ALS3 alleles in each clade. Clade sizes have been equalized for comparison.](image)

**Table 1. ALS3 alleles displayed by clade and tandem repeat copy number**

<table>
<thead>
<tr>
<th>Clade</th>
<th>No. of alleles analysed</th>
<th>Percentage of alleles in each tandem repeat copy number group</th>
<th>Mean no. of repeat copies per allele, ±SD*</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>102</td>
<td>0 0 0 0 1 44 2 2 47 0 0 1 1 0 1</td>
<td>11.7 ± 1.8</td>
</tr>
<tr>
<td>II</td>
<td>94</td>
<td>0 1 1 3 7 12 34 9 29 4 1 0 0 0 0</td>
<td>11.4 ± 1.6</td>
</tr>
<tr>
<td>III</td>
<td>80</td>
<td>1 0 0 4 9 35 23 6 21 0 1 0 0 0 0</td>
<td>10.8 ± 1.6</td>
</tr>
<tr>
<td>E</td>
<td>44</td>
<td>0 2 0 2 0 68 5 23 0 0 0 0 0 0 0</td>
<td>11.4 ± 1.1</td>
</tr>
<tr>
<td>SA</td>
<td>72</td>
<td>0 0 0 1 17 35 6 1 8 4 28 0 0 0 0</td>
<td>11.7 ± 2.5</td>
</tr>
<tr>
<td>Total</td>
<td>392</td>
<td>0.3 0.8 1.8 7.1 27.9 21.9 4.6 27.6 1.8 5.6 0.3 0.3 0.0 0.3</td>
<td>11.4 ± 1.8</td>
</tr>
</tbody>
</table>

*Using the Kolmogorov–Smirnov test, the P values between clades in the relative distribution of the number of tandem repeats per allele were the following: I vs II, P = 0.011; I vs III, P < 0.001; I vs E, P < 0.001; I vs SA, P = 0.001; II vs III, P = 0.001; II vs E, not significant; II vs SA, P = 0.001; III vs E, P < 0.001; III vs SA, P = 0.001; E vs SA, P < 0.001.
**Table 2. Allelic differences within clades**

<table>
<thead>
<tr>
<th>Clade</th>
<th>No. of strains analysed</th>
<th>Percentage of heterozygous strains</th>
<th>Percentage of homoyzogous strains</th>
<th>Mean difference in no. of repeat copies, ± SD*</th>
<th>Most common allelic combination</th>
<th>Percentage strains with 10/13 allelic combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>51</td>
<td>94</td>
<td>6</td>
<td>3.1 ± 1.2</td>
<td>10/13 (82 %)</td>
<td>82</td>
</tr>
<tr>
<td>II</td>
<td>47</td>
<td>83</td>
<td>17</td>
<td>2.0 ± 1.4</td>
<td>11/13 (34 %)</td>
<td>4</td>
</tr>
<tr>
<td>III</td>
<td>40</td>
<td>82</td>
<td>18</td>
<td>2.2 ± 1.5</td>
<td>10/13 (35 %)</td>
<td>35</td>
</tr>
<tr>
<td>E</td>
<td>22</td>
<td>64</td>
<td>36</td>
<td>1.3 ± 1.1</td>
<td>11/13 (45 %)</td>
<td>0</td>
</tr>
<tr>
<td>SA</td>
<td>36</td>
<td>86</td>
<td>14</td>
<td>3.7 ± 2.2</td>
<td>10/15 (33 %)</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>196</td>
<td>84</td>
<td>16</td>
<td>2.6 ± 1.7</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

*The following P values were computed between clades for the relative difference in distribution of the difference in number of tandem repeat copies of a strain using the Kolmogorov–Smirnov test: I vs II, P < 0.001; I vs III, P < 0.001; I vs E, P < 0.001; I vs SA, P = 0.001; II vs III, not significant; II vs E, P = 0.006; II vs SA, P < 0.001; III vs E, P = 0.002; III vs SA, P < 0.001; E vs SA, P < 0.001. The following Fst statistics were computed among clades to test their differentiation: I vs II, θ = 0.149, P < 10^-4; I vs III, θ = 0.077, P < 10^-4; I vs E, θ = 0.376, P < 10^-4; I vs SA, θ = 0.157, P < 10^-4; II vs III, θ = 0.036, P = 7 × 10^-4; II vs E, θ = 0.079, P < 10^-4; II vs SA, θ = 0.135, P < 10^-4; III vs E, θ = 0.188, P < 10^-4; III vs SA, θ = 0.064, P < 10^-4; E vs SA, θ = 0.314, P < 10^-4.*

also unique in that it contained significant proportions of alleles with 9 and 15 tandem repeat copies (Table 1, Fig. 4).

Statistical comparisons between the clades were also made, based on the relative distribution of tandem repeat copies per allele (Table 1). These comparisons revealed that the relative distribution of tandem repeat copies per allele differed significantly between every pair of clades except II and E. Therefore, although clades II and III appeared quite similar with respect to the most prevalent alleles, they were still significantly different with respect to the relative distribution of allelic sizes.

**Clade differences in allelic pairing**

The pairing of ALS3 alleles in strain SC5314 matched one allele with strong adhesive function (12 tandem repeat copies) with another that did not contribute substantially to adhesion in the cell monolayer assays (9 tandem repeat copies; Fig. 2). Because of this allelic pairing, genotyping data were analysed to determine how ALS3 alleles are paired in each clade. In the 196 strains analysed, the majority of strains (84 %) were heterozygous for ALS3 alleles; this trend was also present within each clade (Table 2). However, in some clades, such as clade I, almost all strains were heterozygous for ALS3 alleles, while in other clades such as E, over one-third of the strains were homozygous. The homozygous strains in clade E encoded 11 copies of the tandem repeat in each ALS3 allele, consistent with allelic frequencies summarized in Table 1.

The mean difference in number of tandem repeat copies between the two ALS3 alleles of a strain was 2.6 ± 1.7 for the entire collection of 196 strains examined. The difference in number of tandem repeat copies between the two ALS3 alleles of a strain differed significantly between clades, except in the case of clades II and III (Table 2). The mean difference varied from a low value of 1.3 ± 1.1 tandem repeat copies in clade E to a high of 3.7 ± 2.2 tandem repeat copies in clade SA (Table 2).

Fst statistics were calculated to determine if ALS3 allelic representation in each clade reflected the genetic divergence shown by Ca3 fingerprinting. Results of this analysis showed strong ALS3 genetic differentiation between each pair of clades (Table 2). Fst statistics were also used to determine whether ALS3 allelic composition of the strains was a function of the geographical origin of the strain. For example, 22 of the 47 clade II strains were from South Africa. The θ value for comparison of the clade II South African population to the clade II non-South African strains was 0.0081 (P = 0.65) showing that the two populations were not differentiated. These data further demonstrated the divergence in ALS3 alleles observed between clades.

The frequency of different allelic combinations in the entire strain collection was also examined. Some combinations were prevalent, such as the pairing of one allele containing 10 repeat copies with another containing 13 repeat copies (denoted 10/13). This allelic pair was found in 82 % of clade I strains and was also the most frequent pair in clade III (35 %). In contrast, the 10/13 combination was almost absent in strains from clades II, E and SA (Table 2). In these clades, the 11/13 combination (for II and E) or the 10/15 combination (for SA) predominated.

Differences in allelic pairing are, in part, a result of the differences in prevalent alleles between the clades. However, the combinations observed are unlikely to be random. To assess the non-randomness of allelic combinations, we tested the total collection and then individual clades for the expected proportion of the 10/10 and 13/13 combinations, using Fisher’s Exact test. The entire pool of 196 analysed strains was found to have a deficit in the 10/10 combination (P = 0.046) and in the 13/13 combination (P = 0.031). When clades were analysed, group I alone showed these
deficits ($P < 0.001$). No clade I isolates with a 10/10 combination were found among the 51 strains tested, even though the expected number was 9-9. Only two clade I isolates with a 13/13 combination were found, even though the expected number was 11-3. In contrast, non-clade I strains contained the expected number of 10/10 and 13/13 isolates, and, therefore, no deficit. We then tested if this type of deficit was restricted to clade I or was true for other alleles in other clades. Analysis of the 11/13 allelic pairing did not reveal non-random distribution of this allelic pair either in the whole population or in specific clades.

A qualitative survey of ALS3 allelic combinations suggested that in individual strains, 'short' and 'long' alleles were often paired in the same strain. To test whether this observation was significant, we grouped the shorter alleles together (fewer than 12 repeat copies) and considered them to represent a single allele that we called 'short'. We also grouped the longer alleles together (12 or more repeat copies) and considered them to represent a single allele called 'long'. Twelve repeat copies were chosen as the boundary based on the results of the functional analysis (Fig. 2) and the distribution of alleles (Fig. 3), which showed a dramatic lack of alleles with 12 repeat copies compared to those with 10, 11 or 13. We found a statistically significant excess of strains with a short and long allelic pair in the total population ($P < 0.001$). The excess was still significant ($P < 0.001$) when the 51 clade I strains were removed from the analysis. This result demonstrated that the significance of long/short allelic pairing was not solely due to the preponderance of 10/13 strains in clade I, but was more of a general phenomenon characteristic of the whole C. albicans population.

**Clades are similar for mean number of tandem repeat copies per allele**

Even though differences existed in predominant alleles and in the relative distribution of alleles, the mean number of repeats per allele was similar among the clades. The range was 10·8 ± 1·6 to 11·7 ± 2·5, a difference of 0·9 repeats (Table 1). The mean number of repeats for the ALS3 alleles of two clades (II, E) was the same as the mean for the alleles of all clades, and the mean number for the ALS3 alleles of two more clades (I, SA) differed from the mean by only 0·3 repeats. The small differences observed in the mean number of repeats per allele was remarkable when one considers the dramatic differences demonstrated between clades in the prevalent ALS3 alleles, and in the distribution of tandem repeat copy numbers, the difference in the number of repeats between the alleles of each isolate and the non-random combinations of alleles in individual strains.

**DISCUSSION**

ALS3 alleles that differ in the number of tandem repeat copies within the central domain of the coding region encoded proteins that gave a different functional readout in a commonly used adhesion assay system. Demonstration of this difference is important because it may affect conclusions about the role of Als3p in C. albicans. For example, studying the larger ALS3 allele from strain SC5314 using these assays would lead to the conclusion that Als3p is a C. albicans adhesin, but studying the smaller ALS3 allele would suggest that the adhesive role of Als3p is marginal at best. Allelic variation also can complicate interpretation of results from restoring a wild-type ALS3 copy to a mutant strain. This construction is expected as a means of demonstrating that mutant strains only contain lesions in the gene of interest. Construction of such a replacement strain with the smaller ALS3 allele would lead to the conclusion that the strain contained collateral damage from the mutagenesis process or, minimally, that gene dosage is important for function of the ALS3 product. These results highlight the importance of understanding ALS allelic diversity at the population level and the need for caution in extrapolating evidence from the study of a single allele to the role of the gene in a more global view of C. albicans strains.

Currently, the functional model for Als proteins is that an adhesive N-terminal domain is projected away from the cell surface by the highly glycosylated tandem repeat and C-terminal domains. This relationship between structure and function has been demonstrated for the Epa1p protein of Candida glabrata (Frieman et al., 2002). With this model, the longer ALS3 alleles would produce proteins that project the N-terminal domain farther from the C. albicans cell surface, presumably lending an advantage in contacting host ligands. However, results from our study showed that C. albicans strains tend to be heterozygous for ALS3 alleles, with strains commonly encoding a longer and shorter allele. The prevalence of this allelic combination suggests that it is advantageous to the organism and may be selected for in nature. The role of the shorter Als3p protein in adhesion is not immediately clear, since changes in sequenced alleles appear to be limited to the tandem repeat domain and C-terminal domain, with the amino acid sequence of the N-terminal domain the same for the encoded proteins. Unless adhesive capacity resides in a domain in addition to the N-terminal domain, adhesion of the proteins produced from shorter alleles will not be less avid than that from larger alleles, but may only result in a less accessible protein on the C. albicans cell surface. In the strains assayed, expression of the two alleles appears to be under the same control, and differential allelic expression has yet to be observed (Hoyer et al., 1998). These results suggest that the role of the protein encoded by the shorter allele may be distinct from adhesion and lead to the possibility that strains containing two shorter alleles may be less adherent and possibly less virulent than those with the long/short allele combination. This hypothesis remains to be tested.

The tendency to find a longer and shorter allele paired in C. albicans strains suggests selection of distinct alleles in
the same strain. This result contrasts with the study of ALS7 allelic variation that showed that alleles in the same strain are more similar to each other than expected by chance (Zhang et al., 2003). From these comparisons emerges the conclusion that distinct selective pressures are placed on the various ALS genes. Comparison of DNA sequences and length of the tandem repeat domain for ALS3 and other ALS genes also supports this conclusion of differing selective pressures within the ALS family. For example, despite the range of changes in tandem repeat copy number between ALS3 alleles, the sequence of the N-terminal domain of the encoded protein remains constant, suggesting that there is selective pressure to maintain it. This result differs from that for Als5p, in which amino acid sequence differences within the N-terminal domain are more frequent than those noted for other C. albicans proteins (Hoyer & Hecht, 2001; Zhao et al., 2003). However, for Als5p, there appears to be strong pressure to maintain the number of tandem repeat copies within the central domain in a range much more narrow than for Als3p (R. Jajko & L. L. Hoyer, unpublished observations). Large differences in amino acid sequence within the N-terminal domain of Als9p have also been found (Zhao et al., 2003), and other ALS genes have a far more broad variation in repeat copy number within the central domain (Lott et al., 1999; Zhang et al., 2003; J. A. Nuessen & L. L. Hoyer, unpublished data). Analysis of each gene across the various clades will define these relationships and provide conclusions regarding Als proteins that can be generalized across the C. albicans strain population.

Our analysis of alleleism of ALS3 based on differences in the tandem repeat copy number in the central domain also highlighted differences between the five major genetic clades of C. albicans (Soll & Pujol, 2003). The clade-specific differences observed were not a function of the geographical origin of the strains examined. We found that although ALS3 alleles contained between 6 and 19 tandem repeat copies, the majority of alleles (78%) had 10, 11 or 13 copies, suggesting that these lengths were selected either based on function or as a result of the mechanism of reorganization (e.g. unequal crossing over) (Pujol et al., 1999).

The distribution of alleles differed significantly amongst clades. For instance, while the great majority of clade I isolates carried 10 and 13 tandem repeat alleles, the majority of clade E isolates carried the 11 tandem repeat allele, the majority of clade II isolates carried the 11 and 13 tandem repeat alleles and the majority of clade SA isolates carried the 10 and 15 tandem repeat alleles. What is significant about these results is not the identity of the prevalent alleles, but rather the fact that each clade differed so markedly in allelic prevalence. These differences between clades suggest that while there is some degree of uniformity of prevalent alleles within clades, there is far less uniformity amongst clades. Recently, Pujol et al. (2004) discovered that natural 5-fluorocytosine (flucytosine)-resistant strains were members of a single clade and the great majority of clade I isolates were less susceptible to 5-fluorocytosine than members of all other clades. Resistance and decreased susceptibility were due to a single allele with a mutation in one nucleotide (Dodgson et al., 2004). The former were homozygous and the latter heterozygous for the allele. The authors interpreted these results to suggest that while recombination occurred between strains within a clade, it was far rarer between strains of different clades. Clade-specific differences in the prevalence of particular ALS3 alleles could be explained similarly. Alternatively, different selective pressures on isolates from different clades may determine the most common clade-specific ALS3 alleles, or clade-specific differences in the mechanisms determining tandem repeat number may exist. Functional analysis of additional ALS3 alleles from a wider population of C. albicans isolates will further our understanding of the importance of tandem repeat copy number and the presence of heterozygous ALS3 alleles within the same strain. The genotyping of C. albicans strains presented here prioritizes alleles for these additional studies. Similar analyses for other ALS genes will indicate how broadly applicable these themes are across the ALS family.

ACKNOWLEDGEMENTS

This research was funded by NIH grants DE14158 to L. L. H. and AI2392 to D. R. S.

REFERENCES


