Published in final edited form as:

Fungal Genet Biol. 2007 December ; 44(12): 1298–1309.

Analysis of *ALS5* and *ALS6* allelic variability in a geographically diverse collection of *Candida albicans* isolates

Xiaomin Zhao^a, Soon-Hwan Oh^a, Robert Jajko^a, Daniel J. Diekema^{b,c}, Michael A. Pfaller^c, Claude Pujol^d, David R. Soll^d, and Lois L. Hoyer^{a,*}

^aDepartment of Pathobiology, University of Illinois at Urbana-Champaign, Urbana, IL 61802, USA

^bDepartment of Internal Medicine, University of Iowa, Iowa City, IA 52242, USA

^cDepartment of Pathology, University of Iowa, Iowa City, IA 52242, USA

^dDepartment of Biological Sciences, University of Iowa, Iowa City, IA 52242, USA

Abstract

The *Candida albicans* ALS (agglutinin-like sequence) gene family encodes eight cell-surface glycoproteins, some of which function in adhesion to host surfaces. ALS genes have a central tandem-repeat-encoding domain comprised entirely of head-to-tail copies of a conserved 108-bp sequence. The number of copies of the tandemly repeated sequence varies between *C. albicans* strains and often between alleles within the same strain. Because ALS alleles can encode different-sized proteins that may have different functional characteristics, defining the range of allelic variability is important. Genomic DNA from *C. albicans* strains representing the major genetic clades was PCR-amplified to determine the number of tandemly repeated sequence copies within the *ALS5* and *ALS6* central domain. *ALS5* alleles had 2 to 10 tandem repeat sequence copies (mean = 4.82 copies) while *ALS6* alleles had 2 to 8 copies (mean = 4.00 copies). Despite this variability, tandem repeat copy number was stable in *C. albicans* strains passaged for 3000 generations. Prevalent alleles and allelic distributions varied among the clades for *ALS5* and *ALS6*. Overall, *ALS6* exhibited less variability than *ALS5*. *ALS5* deletions can occur naturally in *C. albicans* via direct repeats flanking the *ALS5* locus. Deletion of both *ALS5* alleles was associated particularly with clades III and SA. *ALS5* exhibited allelic polymorphisms in the coding region 5' of the tandem repeats; some alleles resembled *ALS1*, suggesting recombination between these contiguous loci. Natural deletion of *ALS5* and the sequence variation within its coding region suggest relaxed selective pressure on this locus, and that Als5p function may be dispensable in *C. albicans* or redundant within the Als family.

1. Introduction

The *Candida albicans* ALS gene family includes eight genes (*ALS1* to *ALS7*, and *ALS9*) that encode large cell-surface glycoproteins (Hoyer et al., 2007). The ALS genes share a similar basic organization consisting of a relatively conserved 5' domain, a central domain of tandemly repeated sequence units, and a 3' domain of relatively variable length and sequence (Hoyer, 2001; Hoyer et al., 2007). Although the tandem repeat units in each ALS gene are 108 bp, the basic sequence of the repeat unit is variable and serves to group ALS genes into three subfamilies consisting of i). *ALS1* to *ALS4*; ii). *ALS5* to *ALS7*; and iii). *ALS9* (Hoyer et al.,

*Corresponding author. Mailing address: Department of Pathobiology, University of Illinois at Champaign-Urbana, 2522, VMSB, 2001 S. Lincoln Avenue, Urbana, IL 61802, USA, Phone: (217) 333-5043., Fax: (217) 244-7421., E-mail: lhoyer@uiuc.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

2007). Deletion of individual ALS genes and phenotypic testing of the resulting mutant *C. albicans* strains demonstrated that *ALS1*, *ALS2*, *ALS3*, *ALS4* and *ALS9* contribute to *C. albicans* adhesion (Fu et al., 2002; Zhao et al., 2004, 2005, 2007). Overexpression of ALS genes in *S. cerevisiae* suggested an adhesive role for *ALS5* and *ALS6* (Gaur and Klotz, 1997; Sheppard et al., 2004), but this role has not been demonstrated in *C. albicans*. Despite what is known about Als protein function, it is still unclear whether the family exists to provide *C. albicans* with different specificities of the same basic function (for example, adhesion to a variety of host surfaces) or to provide *C. albicans* with redundancy of critical functions. The high level of allelic variability found within the ALS family complicates studies that address these questions.

ALS allelic variability is most obvious within the central tandem repeat domain of each gene (Hoyer et al., 2007). For a few of the ALS genes, sequence variability exists within the 5' domain, which is believed to encode the main adhesive domain of the Als protein (Hoyer and Hecht, 2000; Zhao et al., 2003), and in repeated regions within the 3' domain, which encodes a serine/threonine-rich, heavily glycosylated portion of the mature protein (Zhang et al., 2003; Zhao et al., 2003). At the *ALS3* locus, *C. albicans* strains tend to maintain heterozygous alleles with respect to the number of copies of the tandem repeat sequence present in the central domain (Oh et al., 2005). In a collection of clinical isolates, the mean difference in number of tandem repeat copies between two *ALS3* alleles in the same strain was 2.6. Phenotypic testing of derivatives of *C. albicans* strain SC5314 showed that the *ALS3* allele with 12 tandem repeat copies made the major contribution to *C. albicans* adhesion to endothelial and epithelial surfaces, while the *ALS3* allele with 9 tandem repeat copies made a significant, but very minor adhesive contribution (Oh et al., 2005). This work suggested the possibility that *C. albicans* maintains two distinct *ALS3* alleles, potentially for different functions. This theme was also illustrated by the study of *ALS9*, which displays the greatest number of types of allelic variability within the ALS family. Within the 5' domain, *ALS9* alleles are 11% different at the nucleotide level (16% different at the amino acid level; Zhao et al., 2003). Like all other ALS genes, *ALS9* tandem repeat copy number varies within the central domain. Within the 3' domain, certain *ALS9* alleles have extra sequence blocks that are absent in other alleles (Zhao et al., 2003). In strain SC5314, the *ALS9-2* allele, but not the *ALS9-1* allele, restored adhesive function to an *als9Δ/als9Δ* strain (Zhao et al., 2007). Examination of the clinical isolate collection indicated extensive recombination at the *ALS9* locus with an obvious preference for *ALS9-2* allelic sequences (Zhao et al., 2003; Zhao et al., 2007).

These examples emphasize the allelic complexity within the ALS family and prompted analysis of other ALS loci to define their allelic variability. Knowledge of allelic variability provides the context required to draw accurate functional conclusions about Als proteins. The focus of this paper is *ALS5* (Gaur and Klotz, 1997; Hoyer and Hecht, 2001) and *ALS6* (Hoyer and Hecht, 2000). *ALS5* and *ALS6* share a cross-hybridizing tandem repeat motif and nearly 100% sequence identity within the 3' domain (Hoyer and Hecht, 2000). The 5' domain of *ALS5* is nearly 80% identical to that of *ALS1* and *ALS3*, while the 5' domain of *ALS6* is relatively unique within the ALS family (Hoyer et al., 2007). This study examines sequence variability within the central tandem repeat domain and also within other regions of the genes that may affect function of its encoded protein. The strain collection that was used for analysis of *ALS3* alleles was also used in this study, which provides the opportunity for direct comparisons between the loci.

2. Materials and methods

2.1. *C. albicans* strains

The collection of clinical isolates used in this study was obtained from three populations previously analyzed by Ca3 fingerprinting (Blignaut et al., 2002; Pujol et al., 1997, 2002) and

included 88 isolates from the United States and Canada, 71 from South Africa, 25 from Europe, 8 from South America, and 4 from Turkey and Israel. Clades in *C. albicans* were originally described by using the Ca3 fingerprinting method (Blignaut et al., 2002; Pujol et al., 1997, 2002; Soll and Pujol, 2003) and have since been confirmed by MultiLocus Sequence Typing (MLST; Robles et al., 2004; Tavanti et al., 2005; Odds et al., 2007; M. E. Bounoux. C. Pujol, D. Diogo, C. Bouchier, D. R. Soll and C. d'Enfert, unpublished data). Clade status of most of the strains used here was confirmed by MLST (Odds et al., 2007; Tavanti et al., 2005). The collection included 51 clade I isolates, 47 clade II isolates, 40 clade III isolates, 22 clade E isolates and 36 clade SA isolates. clade E was subdivided into the two Ca3-defined subgroups (Pujol et al., 2002), Ea (12 strains in this study) and Eb (10 strains) that were recently shown to represent distinct clades by MLST analysis (Tavanti et al., 2005). An additional 1047 *C. albicans* isolates were assembled from those collected by the ARTEMIS Global Antifungal Surveillance Program (Pfaller et al., 2005). Clade assignments were not determined for this collection with the exception of *ALS5*-negative strains that were typed by the Ca3 method. Strains were screened for the presence of *ALS5* using PCR- and Southern blot-based methods (see below). Additional verification that the *ALS5*-negative strains were *C. albicans* was provided by PCR with the LSUF and LSUR primers (Table 1; Boucher et al., 1996), by the PCR-RFLP method using primers ITS1 and ITS4 (Table 1; Mirhendi et al., 2005), and by Ca3 fingerprinting (Pujol et al., 2002).

Tandem repeat copy number stability in ALS coding regions was assessed using four different *C. albicans* strains that were grown for 3000 generations (Pujol et al., 1999). To grow strains for 3000 generations, a single colony was inoculated into a flask of culture medium and incubated with shaking at 37°C. Every 200 generations, the culture was plated, a single colony isolated, and inoculated into fresh culture medium. Strains were stored on agar slants and eventually in glycerol at -80°C.

2.2. ALS tandem repeat genotyping reactions

Genomic DNA was extracted from each *C. albicans* isolate as described previously (Oh et al., 2005) and the size of the tandem repeat domain in each ALS allele determined by PCR using two independent primer pairs (Fig. 1, Table 1). Use of two primer pairs provided an additional control for the accuracy of the results. For *ALS5*, primer pairs were *ALS5RepF/ALS5RepR* and *ALS5Geno2F/ALS5Geno2R*. For *ALS6*, primer pairs were *ALS6GenoF/ALS6/7GenoR* and *ALS6Geno2F/ALS6Geno3R*. PCR reactions were described previously (Oh et al., 2005). Each used 1x Invitrogen Taq polymerase buffer with 1 mM MgCl₂. PCR products were separated on 3.5% polyacrylamide/Tris Borate EDTA gels and stained with ethidium bromide. Amplicon sizes were determined by comparison to the 1 kb ladder (Invitrogen) and to *ALS5* or *ALS6* amplification products from strain SC5314 genomic DNA (Fig. 1). DNA sequencing of alleles from strain SC5314 showed two different *ALS5* alleles with either 4 (GenBank accession number AY227439) or 5 (GenBank accession number AY227440) copies of the tandem repeat sequence. Within the central domain, both alleles of *ALS6* in strain SC5314 have 4 copies of the tandemly repeated sequence (Fig. 1). The sequence of one allele is deposited in GenBank (accession number AY225310). Stability of the tandem repeat domain was assessed using *C. albicans* strains that were passaged serially in vitro for 3000 generations (Pujol et al., 1999). Genomic DNA from generation zero and generation 3000 was amplified by PCR to determine the number of copies of the tandem repeat sequence as described above for *ALS5* and *ALS6*. *ALS3* tandem repeat domain analysis used primer pairs *ALS3GenoF/ALS3GenoR* and *ALS3GenoF2/ALS3GenoR2* (Table 1) as described by Oh *et al.* (2005).

2.3. DNA sequencing of the ALS5 deletion site

The *ALS5* deletion site was defined by PCR amplification with primers upstream and downstream of the *ALS5* coding region and genomic DNA from *C. albicans* strain 20535.027.

Primers ALS5dn2.5kF and ALS5dn2.5kR (Table 1) yielded a product from genomic DNA from the *ALS5*-negative strains, suggesting that the region 2.5 kb 3' of *ALS5* was still present. The same result was obtained for primer pair ALS5up5.3kF and ALS5up5.3kR (Table 1) suggesting that sequences approximately 5.3 kb upstream of *ALS5* were still present. A PCR product was not obtained with primers ALS5up3kF and ALS5up3kR (Table 1) suggesting that the region 3 kb upstream of *ALS5* was deleted in the *ALS5*-negative strains. This information suggested that primers ALS519upF and ALS5dn2.5kR should produce a PCR product. A 2.2 kb product was obtained, suggesting that greater than 8.5 kb of sequence was deleted in the *ALS5*-negative strain. The DNA sequence of this PCR product was compared to information on the *Candida* Genome Database (www.candidagenome.org) to define the deleted region. To determine whether the deletion site was similar in all *ALS5*-negative strains in the collection, genomic DNA from the strains was PCR amplified with three different primer pairs: ALS519upF/5GapR1 which produces a product of 250 bp in strain 20535.027, 5GapF2/5GapR2 which amplifies a 196-bp product, and ALS519upF/5GapF2 which amplifies a 396-bp product (Table 1). PCR products were analyzed on acrylamide gels and their size compared to a standard fragment amplified from strain 20535.027. Strains that are hemizygous for *ALS5* were identified by PCR screening of strains that appeared homozygous for tandem repeat copy number. Strains that yielded a PCR product with primers ALS519upF/ALS5dn2.5R had one chromosome from which *ALS5* was deleted naturally and one chromosome with an intact *ALS5* allele.

DNA sequencing of the region upstream of the *ALS5* locus in clade I strains was complicated by the fact that the primer pairs used for strain 20535.027 (described above) produced multiple PCR products. The periodicity of the products suggested that the 5' region flanking the *ALS5* locus may contain tandem copies of a complex repeated sequence. Primer 5StartR (Table 1), which hybridizes to sequences within the *ALS5* open reading frame, was paired with primer ALS519upF and a 3.8 kb PCR product cloned into pCRBlunt (Invitrogen). The sequence downstream of *ALS5* was amplified using primers ALS5endF (which hybridizes to sequences at the 3' end of the *ALS5* open reading frame) and ALS5dn1.8kR (which hybridizes to sequences 1.8 kb downstream of the *ALS5* stop codon; Table 1). The PCR product was cloned as described above. DNA sequencing was completed at the Roy J. Carver Biotechnology Center (Urbana, IL).

2.4. DNA sequencing of the *ALS5* and *ALS6* coding regions

ALS5 or *ALS6* fragments were amplified by PCR for DNA sequence analysis to assess sequence divergence within the 5' domain. All of these PCR reactions used *Pfu* proofreading polymerase (Stratagene). Primers for amplification of the *ALS5* fragment were ALS5Xho and RTALS5RALT (Table 1), which amplified from nt 1 to 868 within the coding region (based on GenBank sequence AY227439). Primers for amplification of the *ALS6* fragment were ALS6XhoF and RTALS6R (Table 1), which amplified from nt 1 to 927 within *ALS6* (based on GenBank sequence AY225310). PCR reactions were described previously (Zhao et al., 2003). Initial genotyping analysis of the *ALS6* tandem repeat domain in many clade SA strains did not yield a PCR product with primers ALS6GenoF and ALS6/7GenoR. To identify the altered DNA sequence in these strains, PCR products were generated within the 5' domain (using primers ALS6GenoSeq3F and ALS6GenoSeq3R) and within the 3' domain (using primers 688B and ALS6R). When detected, DNA sequence polymorphisms were verified by sequencing on both strands.

2.5. Southern blots

Southern blotting methods were described previously (Hoyer et al., 1995) and used the Genius non-radioactive labeling system (Roche). Probes included an oligonucleotide specific for *ALS5* (Hoyer and Hecht, 2001), a PCR fragment from the 3' domain of *ALS5* that cross-

hybridizes with *ALS6* (Hoyer and Hecht, 2001) and a 215-bp PCR product specific for *ALS6* (Hoyer and Hecht, 2000).

2.6. 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 normally distributed. 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. The same strategy was used to obtain the difference in number of repeat copies per strain expected if combinations of alleles were random. Fst statistics were assessed using Multilocus 1.3 software package, available at <http://www.agapow.net/software/multilocus> (Agapow and Burt, 2001). The program calculates Weir's (Weir, 1996) formulation of Wright's Fst. The calculated statistic, θ , 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), where an Fst of 1 means that the two populations are completely isolated genetically (they have nothing in common). In our study, Fst statistics were used to test if *ALS5* and *ALS6* gene frequencies reflected the genetic divergence shown by DNA fingerprinting among *C. albicans* clades.

3. Results and Discussion

3.1. *ALS5* allelic diversity within the tandem repeat domain

The greatest amount of sequence divergence observed among *ALS* alleles occurs within the central tandem-repeat-encoding domain, mainly due to variability in the number of tandem repeat copies present (Hoyer et al., 2007). For this reason, much of the analysis of allelic variability presented here focused on that region. Within the central domain, *ALS5* alleles in the strains examined encoded between 2 and 10 copies of the tandemly repeated 108-bp sequence (Table 2). The mean number of repeat copies per allele was 4.82 ± 1.61 . Alleles containing 4 tandem repeat copies were the most common and also represented the median of the population. Overall, the clades differed with respect to prevalent *ALS5* alleles and allele distribution (Table 2). Alleles with 4 tandem repeat copies were common in all clades except Eb where alleles with 6 repeat copies represented 88% of the total. Other less common alleles were relatively restricted to specific clades, such as 2 repeat copies in clade III, 5 in clade I, and 8 in clade II and Ea. The relative distribution of the number of repeat copies per allele was significantly different between all clades except III vs. Ea, III vs. SA, and Ea vs. SA (Table 2).

3.2. Spontaneous deletion of *ALS5* in *C. albicans* clinical isolates

Some *C. albicans* strains did not produce PCR products with either *ALS5* tandem repeat genotyping primer pair in this analysis suggesting large sequence polymorphisms at the priming sites or absence of the *ALS5* locus. These strains were further studied by Southern blotting using *ALS5*-specific sequences as probes. Fourteen of the strains from the current collection (7.4%) did not have an *ALS5*-hybridizing band with either the oligonucleotide or PCR product probe (see Materials and methods), suggesting that these strains lacked the *ALS5* locus. Previous work with a collection of 50 clinical isolates yielded 8% *ALS5*-negative strains (Hoyer and Hecht, 2001). Analysis of strains was expanded to a collection of 1047 *C. albicans* isolates (see Methods). In this collection, 17 *ALS5*-negative isolates (1.6%) were identified. Eight of the 17 strains were bloodstream isolates with the other strains isolated from the respiratory tract (3 isolates), abscess or tissue (3), pleural fluid (1), or ascitic fluid (1). These

ALS5-negative strains were isolated from patients ranging in age from infancy (< 1 year of age) to elderly (> 80 years). Variations in the estimates of percent *ALS5*-negative isolates are likely the results of different clade representation within the various strain collections assayed (see below).

Deletion of *ALS5* is due to direct repeats that are present 5' and 3' of the coding region (Fig. 2). The boundaries of the deleted region were estimated by PCR and then pinpointed by DNA sequencing (see Materials and methods). Because of the size differences between *ALS5* alleles, the length of the deleted region varies in each *C. albicans* isolate. However, PCR amplification using primers that span the deleted region (see Materials and methods) produced the same-sized products in each *ALS5*-negative strain suggesting that the deletion occurred in the same general location in all *ALS5*-negative isolates, regardless of their source. PCR amplification of genomic DNA from each of the *ALS5*-negative strains using a primer set specific for each of the other ALS genes (Green et al., 2004) suggested that the remainder of the ALS family is present in each strain.

Identification of *C. albicans* clinical isolates in which both *ALS5* alleles are deleted suggests that there is not a strong selective pressure to maintain this gene. Creation of a viable *als5A/als5A* mutant strain supports the conclusion that *ALS5* is not essential (Zhao et al., in press). The ability of *C. albicans* to lose *ALS5* spontaneously might also suggest that Als5p function is at least partially redundant. It is possible that some strains lose a single copy of *ALS5* without consequence. Such hemizygous strains would be recorded as homozygous using the genotyping strategy presented above. Therefore, the strain collection was screened again using PCR to identify hemizygous *C. albicans* strains (Table 3). Allelic frequencies recorded in Table 1 were adjusted to reflect this new information. Overall, 38% of the strains tested were missing one *ALS5* allele (Table 3). These strains were frequent in all clades except clade I (4% hemizygous). Fisher's exact test showed that the proportion of clade I strains with at least one deleted *ALS5* allele is significantly lower than in the remainder of the strain collection ($P = 4 \times 10^{-13}$). Similarly, the proportion of *ALS5*-deleted clade I alleles was significantly lower than in the remaining strains ($P = 3 \times 10^{-13}$). These results suggest the potential for selective pressure to maintain two functional *ALS5* alleles in clade I strains. Alternatively, clade I may be more recently evolved and has not accumulated as many deleted *ALS5* alleles. Strains with both *ALS5* alleles deleted were found only in clades III and SA (Table 3). Despite the large number of hemizygous strains in clade II, no deleted strains were observed. Among the 43 clade II isolates analyzed, none were lacking both *ALS5* alleles although 5 null strains were expected if allelic combinations were random ($P = 0.019$). This result suggests that selective pressure in clade II strains may maintain at least one functional *ALS5* allele. Fst statistics were used to test if *ALS5* allele composition reflected clade genetic divergence and showed significant genetic differentiation between most pairs of clades (Table 3). These results indicate that in addition to allelic divergence among clades, differences in frequency of deleted alleles were also present. Deleted alleles were found in all clades and appeared to be due to the same recombination mechanism. In contrast, the proportion of deleted alleles (clade I) or the frequency of *ALS5* null strains (clade II) varied dramatically in function of clades. This observation demonstrates that *ALS5* genetic diversity is strongly affected by genetic background and should be taken into consideration in future studies.

3.3. *ALS5* allelic diversity in the 5' domain

Previous work showed the presence of allelic variability within the *ALS5* 5' domain (Hoyer and Hecht, 2001). Subsequent analysis indicated that sequences from strains 1161 and CA1 represented the extremes of this DNA sequence variability (Zhao et al., 2003). Although *ALS5* alleles from other strains tended to be more like the CA1 allele, strains that encoded 1161-like alleles were found in clades III, Ea and Eb (Table 4). However, not all strains in

these clades encoded the sequence variations. The 1161-like *ALS5* sequence variations resembled the sequence of *ALS1* in this region suggesting the possibility of recombination between the highly similar 5' domain sequences of the contiguous *ALS5* and *ALS1* coding regions (Fig. 2; Zhao et al., 2003).

3.4. *ALS6* allelic diversity within the tandem repeat domain

ALS6 alleles encoded between 2 and 8 copies of the tandem repeat sequence within the central domain of the gene (Table 5). The mean number of tandem repeat copies per allele was 4.00 ± 0.98 . Like *ALS5*, the median and most common alleles had 4 tandem repeat copies. Alleles with 3 or 4 tandem repeat copies were predominant in all clades except SA, where alleles with 6 repeat copies represented 65% of those tested. In clades I and Eb, nearly all alleles encoded 4 repeat copies. Despite this similarity in allelic composition, the relative distribution of the number of tandem repeat copies per allele differed significantly between all clades with the exception of II vs. III (Table 5).

Although the overall analysis showed a minority of strains were heterozygous with respect to repeat copy number at the *ALS6* locus (36%), great differences were observed among clades (Table 6). *ALS6*-homozygous strains were more prevalent in clades I, Eb and SA (Table 6). Similar to *ALS5*, the most common allelic combination was 4/4, but the 3/4 combination was observed almost as frequently (Table 6). The mean difference in number of repeat copies between the two alleles of the same strain was less than 1.0 for each clade and varied from 0 in clade Eb to 0.92 ± 0.29 in clade Ea. The mean difference for the entire strain collection was 0.38 ± 0.54 . The mean difference in number of tandem repeats differed significantly between most clades except II, III and Ea (Table 6). Fst statistics were used to test if *ALS6* allele composition reflected clade genetic divergence and showed significant *ALS6* genetic differentiation between most pairs of clades (Table 6).

3.5. *ALS6* allelic diversity outside of the tandem repeat domain

The majority of *ALS6* alleles with 6 tandem repeat copies were found in homozygous strains from clade SA (Table 6) and were rare in other clades (Table 5). The clade SA alleles with 6 tandem repeat copies also were notable in that many failed to produce a PCR product with primers *ALS6*GenoF and *ALS6*/7GenoR. However, Southern blotting showed the presence of a fragment that hybridized to an *ALS6*-specific probe, suggesting that *ALS6* was present (data not shown). PCR products were amplified within the 5' and 3' domains near the sites of genotyping primer hybridization to detect sequence changes that may interfere with amplification of the tandem repeat domain (see Materials and methods). DNA sequence analysis of these PCR products indicated that the 5' domain sequence was like other *ALS6* alleles, but that there was an 18-bp deletion in the 3' domain from nt 1885 to 1902 (numbered using GenBank sequence AY225310 as a reference). This deletion, which was unique to strains in clade SA, prevented annealing of primer *ALS6*/7GenoR and explained the negative PCR reactions. The 18-bp deletion maintains the *ALS6* open reading frame and predicts a protein with a C-terminal domain almost as large as those encoded by other alleles. Because of the heavy glycosylation on this portion of the Als protein (Kapteyn et al., 2000), it is likely that loss of 6 amino acids does not have a great effect on Als6p function.

To determine whether *ALS6* had 5' domain sequence polymorphisms similar to those found in *ALS5*, strains were selected from each clade for DNA sequence analysis. Strains analyzed included P52073 (tandem repeat genotype 4/4) and K713 (3/4) from clade I, P22092 (3/3) and P22095 (3/4) from clade II, P22061 (3/4) and P52098 (3/3) from clade III, P22059 (3/4) from clade Ea and P52013 (4/4) from clade Eb, and G1 (4/4), P69 (3/4), P22073 (4/4), GC58 (6/6) and OKP109 (6/6) from clade SA. Of the clade SA strains, P22073, GC58 and OKP109 contained the 18-bp deletion within the 3' domain of the *ALS6* coding region. PCR primers

were designed to amplify the first 927 nt of the *ALS6* coding region (see Methods). Sequences from strains SC5314 (GenBank accession number AY225310) and 1161 (GenBank accession number AF075293) were included for comparison. Alignment of the resulting sequences showed only three positions where nucleotide changes occurred (nt 358, 414 and 474, numbered according to AY225310). In strains P52073 (clade I) and G1 (clade SA), one of the alleles had a T/G transversion that altered amino acid 120 from Ser to Ala. At position 414, one of the alleles of P52013 (clade Eb) and both alleles of P22073, GC58 and OKP109 (all clade SA with the 3' deletion) had a C/T transition that did not affect the amino acid sequence. The T/C transition in these same strains at position 474 was also silent.

Data from analysis of *ALS6* provides a contrast to those from *ALS5*. While strains with natural *ALS5* deletions were detected among the isolates assayed, each of the isolates encoded *ALS6*. Also, the sequence of the *ALS6* 5' domain was more stable than that of *ALS5*, possibly because the *ALS6* 5' domain sequence is less conserved with other *ALS* 5' domains. In comparison, the 5' domain of *ALS5* is 86% identical to the 5' domain of *ALS1*, which is contiguous on chromosome 6 (Fig. 2). The proximity of *ALS5* and *ALS1* may provide greater opportunity for recombination between them.

3.6. Analysis of *ALS* allelic combinations

Comparisons of data from the current study of *ALS5* and *ALS6* tandem repeat allelic variation to that from analysis of *ALS3* (Oh et al., 2005) shows that the range of alleles was narrower for *ALS5* (2 to 10 tandem repeat copies) and *ALS6* (2 to 8 tandem repeat copies) compared to *ALS3* (6 to 19 tandem repeat copies). The mean number of tandem repeat copies in *ALS3* (11.4 copies) was far larger than for *ALS5* (4.82 copies) or *ALS6* (4.00 copies). Within the same *C. albicans* strain, *ALS3* alleles have a mean of 2.56 tandem repeat copies difference between the two alleles, while the *ALS5* alleles (0.80 copies difference) and *ALS6* alleles (0.38 copies difference) were much more closely matched in size (Table 7). Based on allele frequencies, we also compared the observed difference in number of repeat copies between the two alleles of a strain with that expected if combination of alleles was random (Table 7). For *ALS3*, the observed difference was significantly greater than the expected value. In contrast, *ALS5* and *ALS6* each had a significantly lower mean difference than expected. These results demonstrate that *ALS6* alleles or *ALS5* alleles present in a strain (with two *ALS5* alleles) are globally more similar to each other with respect to tandem repeat copy number than expected by chance. This result is similar to the conclusion for analysis of *ALS7* (Zhang et al., 2003), but opposite to observations for *ALS3* (Table 7; Oh et al., 2005).

In previous work, gene frequency analysis of the tandem repeat domain from *ALS3* showed significant deficits of some *ALS3*-homozygous genotypes and more globally, the preferential pairing of alleles with less than 12 repeat copies with alleles containing more than 12 repeat copies (Oh et al., 2005). In that study, 84% of strains tested were heterozygous for *ALS3* repeat copy number. By comparison, the *ALS5* and *ALS6* loci showed a majority of strains homozygous at these loci (Tables 3 and 6), with only 18% (*ALS5*) and 36% (*ALS6*) of strains having a heterozygous genotype. Overall *ALS5* and *ALS6* presented significant excesses of homozygotes and *ALS3* displayed a deficit of homozygotes that was, nevertheless, non-significant (Table 7). While these global comparisons indicate that allelic distributions in *ALS5* and *ALS6* were distinct from *ALS3*, they can hardly summarize the diversity encountered between clades for these genes. In contrast to *ALS3* where heterozygosity was rather homogeneous among clades (Oh et al., 2005), the percentage of heterozygous strains for both *ALS5* and *ALS6* differed widely in function of the clades, as did the frequency of deleted *ALS5* loci (Tables 3 and 6). Intraclade allelic pairing of tandem repeat copy number in *ALS5* did not significantly differ from the numbers expected if allelic combinations were random. For *ALS6*, clade SA strains presented a highly significant deficit in 6/x heterozygotes (where

x represents any *ALS6* allele with a repeat copy number other than 6; $P = 2.1 \times 10^{-4}$). Altogether, our data indicate that alleles of the various *ALS* genes have unique distributions within the major genetic clades of *C. albicans*, and that *ALS3* allelic combinations are distinct.

3.7. Stability of the tandem repeat domain over serial passage

Analysis of *ALS3* (Oh et al., 2005), *ALS5* and *ALS6* demonstrates the variability present within the tandem repeat domain. Tandem repeat instability has been examined in detail for *Saccharomyces cerevisiae FLO1*, which is part of a gene family encoding cell surface proteins involved in yeast flocculation (Verstrepen et al., 2004, 2005; Verstrepen and Klis, 2006). This work showed that frequent recombination events between intragenic tandem repeat sequences results in readily detected expansion and contraction of gene size, creating quantitative alterations in adhesion phenotypes (Verstrepen et al. 2005). The work in *S. cerevisiae* was suggested to serve as a model for how tandemly repeated sequences contribute to plasticity of adhesin-encoding genes in pathogenic fungi. Previous Southern-blot-based attempts to detect changes in *ALS* gene tandem repeat copy number over 530 generations of serial passage of *C. albicans* strains in vitro did not detect any changes (Hoyer et al., 1995). These results suggested that the *ALS* tandem repeats are relatively stable compared to other types of repeated DNA in the *C. albicans* genome (Hoyer et al., 1995). This analysis was revisited using PCR genotyping in order to detect smaller changes that might not be apparent in the context of larger restriction fragments, and to extend analysis over a greater number of generations of passage. Tandem repeat genotyping primers for *ALS3*, *ALS5* and *ALS6* (Table 1) were used to amplify genomic DNA from four *C. albicans* strains that had been grown for 3000 generations in vitro (Fig. 3; Pujol et al., 1999). The size of the tandem repeat region for each gene in each strain did not change, consistent with the conclusion that, despite the diversity of *ALS* alleles present among *C. albicans* clinical isolates, the number of tandem repeat copies present in a given allele is relatively stable. Additional investigation is required to determine if copy number changes would be more obvious for *C. albicans* strains passaged under different environmental conditions, such as in vivo.

Acknowledgements

We thank Richard Hollis and Lauren Wrobel for their work with the collection of *C. albicans* clinical isolates. This research was funded by NIH grants DE14158 to L.L.H. from the National Institute of Dental and Craniofacial Research and AI2392 to D.R.S. from the National Institute of Allergy and Infectious Disease, National Institutes of Health. The ARTEMIS Global Antifungal Surveillance Program is supported by a research grant from Pfizer Inc. to M.A.P. and D.J.D. This investigation was conducted in a facility constructed with support from Research Facilities Improvement Program Grant Number C06 RR16515-01 from the National Center for Research Resources, National Institutes of Health.

References

- Agapow P-M, Burt A. Indices of multilocus linkage disequilibrium. *Mol Ecol Notes* 2001;1:101–102.
- Blignaut E, Pujol C, Lockhart S, Joly S, Soll DR. Ca3 fingerprinting of *Candida albicans* isolates from human immunodeficiency virus-positive and healthy individuals reveals a new clade in South Africa. *J Clin Microbiol* 2002;40:826–836. [PubMed: 11880401]
- Boucher H, Mercure S, Montplaisir S, Lemay G. A novel group I intron in *Candida dubliniensis* is homologous to a *Candida albicans* intron. *Gene* 1996;180:189–196. [PubMed: 8973366]
- Fu Y, Ibrahim AS, Sheppard DC, Chen YC, French SW, Cutler JE, Filler SG, Edwards JE Jr. *Candida albicans* Als1p: an adhesin that is a downstream effector of the *EFG1* filamentation pathway. *Mol Microbiol* 2002;44:61–72. [PubMed: 11967069]
- Gaur NK, Klotz SA. Expression, cloning, and characterization of a *Candida albicans* gene, *ALA1*, that confers adherence properties upon *Saccharomyces cerevisiae* for extracellular matrix proteins. *Infect Immun* 1997;65:5289–5294. [PubMed: 9393828]

- Green CB, Cheng G, Chandra J, Mukherjee P, Ghannoum MA, Hoyer LL. RT-PCR detection of *Candida albicans* ALS gene expression in the reconstituted human epithelium (RHE) model of oral candidiasis and in model biofilms. *Microbiology* 2004;150:267–275. [PubMed: 14766904]
- Hoyer LL. The ALS gene family of *Candida albicans*. *Trends Microbiol* 2001;9:176–180. [PubMed: 11286882]
- Hoyer LL, Scherer S, Shatzman AR, Livi GP. *Candida albicans* ALS1: domains related to a *Saccharomyces cerevisiae* sexual agglutinin separated by a repeating motif. *Mol Microbiol* 1995;15:39–54. [PubMed: 7752895]
- Hoyer LL, Payne TL, Hecht JE. Identification of *Candida albicans* ALS2 and ALS4 and localization of Als proteins to the fungal cell surface. *J Bacteriol* 1998;180:5334–5343. [PubMed: 9765564]
- Hoyer LL, Hecht JE. The ALS6 and ALS7 genes of *Candida albicans*. *Yeast* 2000;16:847–855. [PubMed: 10861907]
- Hoyer LL, Hecht JE. The ALS5 gene of *Candida albicans* and analysis of the Als5p N-terminal domain. *Yeast* 2001;18:49–60. [PubMed: 11124701]
- Hoyer LL, Green CB, Oh S-H, Zhao X. Discovering the secrets of the *Candida albicans* agglutinin-like sequence (ALS) gene family—a sticky pursuit. *Med Mycol*. 2007in press
- Kapteyn JC, Hoyer LL, Hecht JE, Muller WH, Andel A, Verkleij AJ, Makarow M, Van Den Ende H, Klis FM. The cell wall architecture of *Candida albicans* wild-type cells and cell wall-defective mutants. *Mol Microbiol* 2000;35:601–611. [PubMed: 10672182]
- Mirhendi H, Makimura K, Zomorodian K, Maeda N, Ohshima T, Yamaguchi H. Differentiation of *Candida albicans* and *Candida dubliniensis* using a single-enzyme PCR-RFLP method. *Jpn J Infect Dis* 2005;58:235–237. [PubMed: 16116258]
- Odds FC, Bougnoux ME, Shaw DJ, Bain JM, Davidson AD, Diogo D, Jacobsen MD, Lecomte M, Li SY, Tavanti A, Maiden MC, Gow NA, d'Enfert C. Molecular phylogenetics of *Candida albicans*. *Eukaryot Cell*. 200710.1128/EC.00041-07
- Oh S-H, Cheng G, Nuessen JA, Jajko R, Yeater KM, Zhao X, Pujol C, Soll DR, Hoyer LL. Functional specificity of *Candida albicans* Als3p protein and clade specificity of ALS3 alleles discriminated by the number of copies of the tandem repeat sequence in the central domain. *Microbiology* 2005;151:673–681. [PubMed: 15758214]
- Pfaller MA, Boyken L, Messer SA, Tendolkar S, Hollis RJ, Diekema DJ. Comparison of results of voriconazole disk diffusion testing for *Candida* species with results from a central reference laboratory in the ARTEMIS global antifungal surveillance program. *J Clin Microbiol* 2005;43:5208–5213. [PubMed: 16207985]
- Pujol C, Joly S, Lockhart SR, Noel S, Tibayrenc M, Soll DR. Parity among the randomly amplified polymorphic DNA method, multilocus enzyme electrophoresis, and Southern blot hybridization with the moderately repetitive DNA probe Ca3 for fingerprinting *Candida albicans*. *J Clin Microbiol* 1997;35:2348–2358. [PubMed: 9276415]
- Pujol C, Joly S, Nolan B, Srikantha T, Soll DR. Microevolutionary changes in *Candida albicans* identified by them complex Ca3 fingerprinting probe involve insertions and deletions of the full-length repetitive sequence RPS at specific genomic sites. *Microbiology* 1999;145:2635–2646. [PubMed: 10537185]
- Pujol C, Pfaller MA, Soll DR. Ca3 fingerprinting of *Candida albicans* bloodstream isolates from the United States, Canada, South America, and Europe reveals a European clade. *J Clin Microbiol* 2002;40:2729–2740. [PubMed: 12149321]
- Robles JC, Koreen L, Park S, Perlin DS. Multilocus sequence typing is a reliable alternative method to DNA fingerprinting for discriminating among strains of *Candida albicans*. *J Clin Microbiol* 2004;42:2480–2488. [PubMed: 15184424]
- Sheppard DC, Yeaman MR, Welch WH, Phan QT, Fu Y, Ibrahim AS, Filler SG, Zhang M, Waring AJ, Edwards JE Jr. Functional and structural diversity in the Als protein family of *Candida albicans*. *J Biol Chem* 2004;279:30480–30489. [PubMed: 15128742]
- Soll DR, Pujol C. *Candida albicans* clades. *FEMS Immunol Med Microbiol* 2003;39:1–7. [PubMed: 14556989]

- Tavanti A, Davidson AD, Fordyce MJ, Gow NA, Maiden MC, Odds FC. Population structure and properties of *Candida albicans*, as determined by multilocus sequence typing. *J Clin Microbiol* 2005;43:5601–5613. [PubMed: 16272493]
- Verstrepen KJ, Klis FM. Flocculation, adhesion and biofilm formation in yeasts. *Mol Microbiol* 2006;60:5–15. [PubMed: 16556216]
- Verstrepen KJ, Reynolds TB, Fink GR. Origins of variation in the fungal cell surface. *Nat Rev* 2004;2:533–540.
- Verstrepen KJ, Jansen A, Lewitter F, Fink GR. Intragenic tandem repeats generate functional variability. *Nat Genet* 2005;37:986–990. [PubMed: 16086015]
- Weir, BS. Genetic data analysis II. Sinauer Associates, Inc; Sunderland, MA, USA: 1996.
- Zhang N, Harrex AL, Holland BR, Fenton LE, Cannon RD, Schmid J. Sixty alleles of the *ALS7* open reading frame in *Candida albicans*: *ALS7* is a hypermutable contingency locus. *Genome Res* 2003;13:2005–2017. [PubMed: 12952872]
- Zhao X, Pujol C, Soll DR, Hoyer LL. Allelic variation in the contiguous loci encoding *Candida albicans* *ALS5*, *ALS1* and *ALS9*. *Microbiology* 2003;149:2947–2960. [PubMed: 14523127]
- Zhao X, Oh S-H, Cheng G, Green CB, Nuessen JA, Yeater KM, Leng RP, Brown AJP, Hoyer LL. *ALS3* and *ALS8* represent a single locus that encodes a *Candida albicans* adhesin; functional comparisons between Als3p and Als1p. *Microbiology* 2004;150:2415–2428. [PubMed: 15256583]
- Zhao X, Oh S-H, Yeater KM, Hoyer LL. Analysis of the *Candida albicans* Als2p and Als4p adhesins suggests the potential for compensatory function within the Als family. *Microbiology* 2005;151:1619–1630. [PubMed: 15870470]
- Zhao X, Oh S-H, Hoyer LL. Unequal contribution of *ALS9* alleles to adhesion between *Candida albicans* and vascular endothelial cells. 2007In press
- Zhao X, Oh S-H, Hoyer LL. Deletion of *ALS5*, *ALS6* or *ALS7* increases adhesion of *Candida albicans* to human vascular endothelial and buccal epithelial cells. 2007In press

Fig. 1.

Schematic of *ALS5*, *ALS6* and *ALS3* from *C. albicans* strain SC5314 showing the 5' domain, central tandem repeat domain, and 3' domain. Genes are drawn to scale. Shading patterns indicate regions of sequence identity between the various genes. Individual tandem repeat units are indicated in the central domain of each gene, with five copies in *ALS5-1*, and 4 copies in *ALS5-2* and in each of the *ALS6* alleles. Although the tandem repeat units are similar across the ALS family, those in *ALS3* (spotted) do not cross-hybridize at high stringency with the tandem repeats from *ALS5* or *ALS6* (cross-hatched; Hoyer et al., 1998). ALS genes can be grouped into subfamilies based on this difference in the sequence of the tandem repeat unit (Hoyer et al., 2007). Arrows (not drawn to scale) mark the approximate location of PCR primers used to determine the number of tandem repeat copies within each allele (see Materials and methods). Allelic sequences for *ALS5* (AY227439, AY227440) and *ALS3* (AY223551, AY223552) from strain SC5314 were deposited in GenBank, but only one *ALS6* allele was reported (AY225310). Sequence AY225310 is identical to one reported by the genome sequencing project (AACQ01000074) except for two nucleotides that result in two amino acid sequence changes (755 S to R and 763 S to F) in the C-terminal domain of the protein. These differences could reflect allelic variation at the *ALS6* locus. *ALS5* is located in *C. albicans* chromosome 6, while *ALS6* is on chromosome 3, and *ALS3* is on chromosome R (Hoyer, 2001). Solid bars indicate Southern blot probes for *ALS5* and *ALS6* (see Materials and methods). The asterisk above the *ALS5* schematic indicates the location of the oligonucleotide probe.

Fig. 2.

(a) Scale drawing of the chromosome 6 region that encodes *ALS5* to indicate the sequences deleted in naturally occurring *ALS5*-negative strains. DNA sequence information was taken from the *Candida* Genome Database (<http://www.candidagenome.org>) Contig 19-20233; sequence coordinates are shown. Deletion of *ALS5* is mediated by direct repeats that are 5' and 3' of *ALS5*. The gap in the line below the schematic chromosome indicates the deleted region, which extends from approximately 3.5 kb upstream of *ALS5* to 1.0 kb 3' of the *ALS5* stop codon. The total size of the deleted fragment depends upon the size of the *ALS5* allele encoded on that chromosome. An *ALS5*-negative strain has undergone recombination events to remove both alleles from the genome. Alignment of the direct repeat sequence from 5' of *ALS5* (top line; coordinates 59020 to 59425) and 3' of *ALS5* (bottom line; coordinates 67632 to 67877), and the sequence derived from an *ALS5*-negative clinical isolate (middle line), is shown. Vertical arrows above the sequence alignment delimit the region in which recombination occurs between the 5' and 3' direct repeats. (b) Same DNA sequences from upstream and downstream of *ALS5* that were shown in (a), but with vertical lines to indicate sequence identity within the region of *ALS5* deletion. Vertical arrows delimit the region used to calculate percentage identity (85.3%) between the sequences. Calculation of the length of the direct repeat sequences is complicated because regions of high identity are interspersed with sequence gaps. (c) DNA sequences from the same region in a clade I isolate (see Materials and methods). The similar percentage identity for these sequences (84.7%) to that for the sequences in (b) suggests that recombination to delete *ALS5* should be able to occur in this isolate. The higher percentage of sequence identity 3' of the region delimited by arrows suggests an even higher recombination potential in this strain. It is unclear why strains in clade I have not accumulated deleted *ALS5* alleles similar to strains in other clades.

Fig. 3.

Ethidium bromide-stained agarose gel of PCR products from amplification of the tandem repeat region of various ALS genes. Products from amplification of strain SC5314 genomic DNA are shown to demonstrate the PCR strategy illustrated in Fig. 1. Primers ALS3GenoF2 and ALS3GenoR2 were used to amplify the tandem repeat domain of *ALS3*, primers ALS5Geno2F and ALS5Geno2R for *ALS5*, and ALS6Geno2F with ALS6Geno3R for *ALS6* (Table 1). The tandem repeat domain from each of these genes was also amplified from genomic DNA extracted from strains RP10.3 and RP5.1 grown for either zero or 3000 generations (labeled “0 gen” and “3000 gen”, respectively in the figure). Co-migration of the PCR products from the strains at generation 0 and 3000 suggests stability of the tandem repeat copy number in each gene over serial passage in vitro. A similar result was obtained from analysis of two other *C. albicans* strains grown for 3000 generations (RP39.1 and RP39.2) and with the second set of PCR genotyping primers for each ALS gene (data not shown). Molecular sizes (in kb) are shown to the left of the figure.

Table 1

Oligonucleotide primers used in this study

Primer name	Primer sequence (5' – 3')
LSUF	ATC AAC TTA GAA CTG GTA CGG
LSUR	GAT AGT AGA TAG GGA CAG TGG
ITS1	TCC GTA GGT GAA CCT GCG G
ITS4	TCC TCC GCT TAT TGA TAT GC
ALS5RepF	TTT CTC CCT CAG ATA ATA ACC AGT ATC AAT
ALS5RepR	AAG ACA GTT CTT CCA ATG GAT CA
ALS5Geno2F	GAC GCT TAT ATT TCT CCC TCA GAT AA
ALS5Geno2R	ATA CTT GAT GAC TGC TCA ACC AGA
ALS6GenoF	GGA TGG CAA AAA GGG AAA TGA T
ALS6/7GenoR	AAC CCA ATT GAG CTT GAT GGA A
ALS6Geno2F	CCT CTT ATA TAC TTT TGG ACA TCA TAC ACA
ALS6Geno3R	GGT GGC AGA CGT ACT GGA CT
ALS3GenoF	ACC TTA CCA TTC GAT CCT AAC C
ALS3GenoR	GAT GGG GAT TGT GAA GTG G
ALS3GenoF2	CCA CAA CAC ATA CTA ATC CAA CTG A
ALS3GenoR2	TGT AGA CCA CAA AGT TGT ATG GTT G
ALS5dn2.5kF	CTG CAA TGA TAT TGA TCT CAT
ALS5dn2.5kR	ATG TAA CCT GTC TAG AGA TGA
ALS5up5.3kF	GTT GTC TCT GCA ATT GTT GTA
ALS5up5.3kR	ACA TTA TAG TGG CAG TGG TTG
ALS5up3kF	GGT ACA ACA TTT TAC TTG TGA CC
ALS5up3kR	GTA GCA CGT AAT CGA TAA TAA C
ALS519upF	TGT CTG AAT GCA GTA TTA GGA GGC
5GapR1	AAT TTT CGT GTA ATC GTG ATC
5GapF2	CAA AGA AAA TAC ACA GAG AAG
5GapR2	TCA TCT TCT ATC GAA CAG TAG
5StartR	AGC ATT GGA CCA AGT TAA TGA GTC
ALS5endF	CCA CTA AAC ATC CTT CCT GGT TGC
ALS5dn1.8kR	CTT AGC AAA GAG CTG TAA AGG
ALS5Xho	CCC CTG GAG ATG ATT CAA CAA TTT ACA TTG TTA TTC C
RTALS5RALT	ATT GAT ACT GGT TAT TAT CTG AGG GAG AAA
ALS6XhoF	CCC CTC GAG ATG AAG ACA GTA ATA CTA TTA CAT C
RTALS6R	ATC ATT TCC CTT TTT GCC ATC C
ALS6GenoSeq3F	CTA TTG CAC GAT AGC AAT GGC
ALS6GenoSeq3R	CTG ATC CAG TCC ATA TAC TAG TG
688B	GAA ACG ATA ACT ACA GGG C
ALS6R	CCC CTC GAG AAA AAC AGA ACA AAA AAA ACG ACA CC

Table 2
ALS5 alleles displayed by clade and tandem repeat copy number

Clade	No. alleles analyzed	Percent alleles in each tandem repeat copy number group									Mean no. of repeat copies per allele ^a
		2	3	4	5	6	7	8	9	10	
I	96	0	0	65	33	0	1	1	0	0	4.43 ± 0.71
II	57	2	2	35	5	3	7	42	2	2	6.16 ± 2.03
III	44	25	0	61	0	0	2	5	0	7	4.16 ± 2.11
Ea	22	0	0	73	9	0	0	18	0	0	4.82 ± 1.56
Eb	17	6	0	6	0	88	0	0	0	0	5.65 ± 1.06
SA	41	2	5	78	10	0	0	0	2	2	4.27 ± 1.30
Total	277	5	1	57	15	6	3	11	1	2	4.82 ± 1.67

^a Mean number of repeat copies per allele ± standard deviation. Frequencies are for ALS5 alleles after strains were tested for hemizygous configuration. 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* < 0.001; I vs. III, *P* < 0.001; I vs. Eb, *P* < 0.001; I vs. SA, *P* < 0.001; II vs. III, *P* < 0.001; II vs. Ea, *P* = 0.006; II vs. Eb, *P* < 0.001; II vs. SA, *P* < 0.001; III vs. Ea, not significant; III vs. Eb, *P* < 0.001; III vs. SA, not significant; Ea vs. Eb, *P* < 0.001; Ea vs. SA, not significant; Eb vs. SA, *P* < 0.001. Bold type is used for alleles that comprise more than 10% of the total alleles studied.

Table 3

ALS5 allelic differences within clades

Clade	Number of strains analyzed	Percent heterozygous strains	Percent heterozygous strains	Percent heterozygous strains ^a	Percent deleted strains ^a	Most common allelic combinations ^b
I	49	43	53	4	0	4/4 (41%) 5/5 (12%) 8/Δ (39%) 4/Δ (12%) 4/Δ (49%)
II	43	12	21	67	0	Δ (15%) 8/8 (17%)
III	39	10	18	56	15	4/Δ (58%) 4/Δ (17%) 6/Δ (60%) 4/Δ (31%) Δ/Δ (23%)
Ea	12	0	83	17	0	4/Δ (38%) 4/5 (12%)
Eb	10	10	60	30	0	
SA	35	6	34	40	23	
Total	188	18	37	38	7	

The following Fst statistics were computed between clades to test *ALS5* genetic differentiation: I vs. II, $\theta = 0.247$, $P < 10^{-4}$; I vs. III, $\theta = 0.243$, $P < 10^{-4}$; I vs. Ea, $\theta = 0.058$, not significant; I vs. Eb, $\theta = 0.506$, $P < 10^{-4}$; I vs. SA, $\theta = 0.183$, $P < 10^{-4}$; II vs. III, $\theta = 0.059$, $P = 0.0003$; II vs. Ea, $\theta = 0.141$, $P = 0.0017$; II vs. Eb, $\theta = 0.313$, $P < 10^{-4}$; II vs. SA, $\theta = 0.079$, $P = 0.0009$; III vs. Ea, $\theta = 0.157$, $P = 0.0001$; III vs. Eb, $\theta = 0.359$, $P < 10^{-4}$; III vs. SA, $\theta = 0.013$, not significant; Ea vs. Eb, $\theta = 0.490$, $P = 0.0001$; Ea vs. SA, $\theta = 0.111$, $P = 0.023$; Eb vs. SA, $\theta = 0.397$, $P < 10^{-4}$.

^a Hemizygous strains: strains where one *ALS5* allele is missing. Deleted strains: strains that are lacking the *ALS5* locus.

^b Genotypes found in more than 10% of the isolates of each sample. Δ, deleted allele.

Table 4
Alignment of *ALS5* 5' domain nucleotide sequences from various alleles

Strain Name	Clade ^a	Nucleotide sequence (nt 324-nt 390) ^b	Number of Tandem Repeat Copies ^c
CA1 ^d	I	TCTTCATTAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	6
SC5314-1 ^d	I	TCCTCATTAATAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	5
SC5314-2 ^d	I	TCTTCATTAATAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	4
K221	I	TCCTCATTAATAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	4.5
P57047	II	TCTTCATTAATAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	4.4
P52013	Eb	TCCTCATTAATAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	4.4
OKP77	SA	TCCTCATTAATAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	6.6
OKP109	SA	TCTTCATTAATAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	4.4
P69	SA	TCCTCATTAATAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	10,10
P22095	II	TCCTCATTAATAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	8.Δ
P57045	Eb	TCTTCATTAATAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	6.6
P52073	I	TCCTCATTAATAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	4.4
P52098	III	TCTTCATTAATAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	4.4
P22059	Ea	TCTTCATTAATAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	4.Δ
P22079	III	TCCTCATTAATAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	2.Δ
1161 ^d		TCTTCATTAATAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	2
<i>ALS1</i>		TCTTCATTAATAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	N/A
Consensus Sequence		TCTTCATTAATAATGTACAGTGAACAATAATTTGAGATCATCTATTAAAGCTTTGGGTACGGTTACTTTA	

^a Clade assignments, derived by Ca3 fingerprinting, are shown. Using this method, strains CA1 and 1161 do not cluster to any of the major clades. These strains are included because they are the source of *ALS5* sequences in GenBank. Other strains chosen for this analysis represent different *ALS5* tandem repeat allelic combinations within the various clades.

^b Nucleotide sequences from each strain in frame. Only the portion of the 5' domain sequence that showed the greatest sequence variability is shown. Nucleotides that do not match the majority consensus sequence are underlined. Double underlining denotes nucleotides that alter the predicted amino acid sequence. The *ALS1* sequence shown is from GenBank accession number L25902.

^c Δ, deleted allele. N/A, not applicable. The distribution of tandem repeat copy numbers is different for *ALS1* and is not applicable to the current study.

^d Nucleotide sequences from these strains are in GenBank. CA1 = accession number AF025429; SC5314-1 = accession number AY227440; SC5314-2 = accession number AY227439; 1161 = accession number AF068866. For these alleles, the number of tandem repeat copies was determined by DNA sequencing of a cloned fragment, so the results for only one allele are shown. For all other entries, results were derived by PCR, so tandem repeat copy number for both alleles is shown.

Table 5
ALS6 alleles displayed by clade and tandem repeat copy number

Clade	No. alleles analyzed	Percent alleles in each tandem repeat copy number group							Mean no. of repeat copies per allele ^a
		2	3	4	5	6	7	8	
I	102	0	2	97	1	0	0	0	3.99 ± 0.17
II	92	1	52	38	2	4	0	2	3.65 ± 1.00
III	74	0	54	38	0	5	3	0	3.64 ± 0.92
Ea	24	0	54	46	0	0	0	0	3.46 ± 0.51
Eb	20	0	10	90	0	0	0	0	3.90 ± 0.31
SA	60	0	8	22	5	65	0	0	5.27 ± 1.07
Total	372	0.3	30	55	1.6	12	0.5	0.5	4.00 ± 0.98

^aMean number of repeat copies per allele ± standard deviation. 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* < 0.001; I vs. III, *P* < 0.001; I vs. Ea, *P* < 0.001; I vs. Eb, *P* < 0.001; I vs. SA, *P* < 0.001; II vs. Ea, *P* = 0.004; II vs. Eb, *P* < 0.003; II vs. SA, *P* < 0.001; III vs. Ea, *P* = 0.008; III vs. Eb, *P* = 0.003; III vs. SA, *P* < 0.001; Ea vs. Eb, *P* = 0.019; Ea vs. SA, *P* < 0.001; Eb vs. SA, *P* < 0.001. Bold type is used for alleles that comprise more than 10% of the total alleles studied.

Table 6

ALS6 allelic differences within clades

Clade	Number of strains analyzed	Percent heterozygous strains	Percent homozygous strains	Mean difference in number of repeat copies ^a	Most common allelic combinations ^b	
I	51	4	96	0.06 ± 0.31	4/4 (96%)	4/4 (11%)
II	46	59	41	0.65 ± 0.64	3/4 (54%)	3/3 (24%)
III	37	57	43	0.55 ± 0.50	3/4 (57%)	3/3 (27%)
Ea	12	92	6	0.92 ± 0.29	3/4 (92%)	4/4 (11%)
Eb	10	0	100	0	4/4 (90%)	
SA	30	20	80	0.20 ± 0.41	6/6 (63%)	3/4 (18%)
Total	186	36	64	0.38 ± 0.54	4/4 (38%), 6/6 (12%)	3/3 (12%)

^aMean ± standard deviation. The following *P* values were computed for the relative difference in distribution of the difference in number of repeat copies between the two alleles of a strain using the Kolmogorov-Smirnov test: I vs. II, *P* < 0.001; I vs. III, *P* < 0.001; I vs. Ea, *P* < 0.001; I vs. Eb, *P* < 0.001; I vs. SA, *P* = 0.025; II vs. III, not significant; II vs. Ea, not significant; II vs. Eb, *P* < 0.001; II vs. SA, *P* = 0.001; III vs. Ea, not significant; III vs. Eb, *P* < 0.001; III vs. SA, *P* = 0.001. The following *F*_{st} statistics were computed between clades to test *ALS6* genetic differentiation: I vs. II, *θ* = 0.491, *P* < 10⁻⁴; I vs. III, *θ* = 0.527, *P* < 10⁻⁴; I vs. Ea, *θ* = 0.648, *P* < 10⁻⁴; I vs. Eb, *θ* = 0.041, not significant; I vs. SA, *θ* = 0.681, *P* < 10⁻⁴; II vs. III, *θ* = -0.010, not significant; II vs. Ea, *θ* = -0.018, not significant; II vs. Eb, *θ* = 0.288, *P* < 10⁻⁴; II vs. SA, *θ* = 0.336, *P* < 10⁻⁴; III vs. Ea, *θ* = -0.019, not significant; III vs. Eb, *θ* = 0.306, *P* = 10⁻⁴; III vs. SA, *θ* = 0.344, *P* < 10⁻⁴; Ea vs. Eb, *θ* = 0.325, *P* = 1.3 × 10⁻³; Ea vs. SA, *θ* = 0.386, *P* < 10⁻⁴; Eb vs. SA, *θ* = 0.491, *P* = 10⁻⁴.

^bGenotypes found in more than 10% of the isolates of each sample.

Table 7

Distinct patterns of non-random ALS allele combinations indicated by difference in the number of tandem repeat copies per strain

ALS gene (number of strains)	Mean difference in number of repeat copies per strain $\pm$ SD		Significance ^c	Number of homozygous strains		Significance ^d
	Observed	Expected ^b		Observed	Expected ^b	
ALS3 ^a (196)	2.56 $\pm$ 1.70	1.97 $\pm$ 1.61	$P < 10^{-3}$	31	41.6	Not significant
ALS5 (188)	0.80 $\pm$ 1.85 ^e	1.59 $\pm$ 1.68 ^e	$P < 10^{-3}$	70 ^f	37.4 ^f	$P < 10^{-3}$
ALS6 (186)	0.38 $\pm$ 0.54	0.95 $\pm$ 1.01	$P < 10^{-3}$	119	75.2	$P < 10^{-3}$

^aData from Oh et al. (2005).

^bExpected under the hypothesis that combination of alleles were random.

^cThe significance was calculated by using the non-parametric Kolmogorov-Smirnov test to compare the distribution of the difference in number of repeat copies between the two alleles of a strain for observed and expected data.

^dThe significance was calculated by using the χ^2 statistic.

^eObserved and expected values were for 101 isolates with two ALS5 alleles. The remaining 87 isolates had lost one or two ALS5 alleles and were not used to compute averages.

^fStrains that were missing one or two ALS5 alleles were not included among ALS5-homozygous genotypes.