Sequence analysis reveals complex mutational processes for allele length variation at two polymorphic microsatellite loci in *Candida albicans*

Paula Sampaio[1], **Alexandra Correia**[1], **Leonor Gusmão**[2], **António Amorim**[2,3] and **Célia Pais**[1*]

[1]CBMA-Centre of Molecular and Environmental Biology, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
[2]IPATIMUP-Institute of Pathology and Molecular Immunology of Porto University, Rua Roberto Frias s/n, 4200-465 Porto, Portugal
[3]Faculty of Sciences, Porto University, Porto, Portugal

Although several microsatellite loci have been described for *Candida albicans* very little is known about the mutational dynamics of microsatellite evolution in this species. In the present work the mutational events that generated allele length variability at two compound microsatellite loci, CAI and CEF3, were studied. Allele sequence polymorphism and structural variability were investigated in a population of 72 *C. albicans* clinical isolates and three molecular mechanisms namely, (i) addition/deletion of microsatellite repeated units, (ii) addition/deletion of mononucleotide units with the formation of secondary structures, and (iii) single base mutations in the flanking regions, contributed to the generation of allelic variation at these loci. Although the molecular events identified were similar for both loci, allele frequencies indicated that patterns driving microsatellite evolution might differ. The addition/deletion of repeated units was the mechanism that most contributed to generate length variability at CAI, while mutations in the flanking regions of CEF3 seem to have contributed equally to allele polymorphism.

Keywords: *Candida albicans*, Microsatellite, Allelic variation, CAI, CEF3.

1. Introduction

Microsatellites are short tandemly repeated (STR) sequence motifs consisting of 1-6 bp, very abundant in eukaryotic genomes, which have been increasingly used as the markers of choice in genetic analysis. These loci present a high degree of variability most likely due to the propensity of slippage on the repeated arrays, the primary mutational mechanism leading to changes in microsatellite length [23, 35, 33, 13, 36]. Studies of microsatellite patterns of mutation indicated that the gains of repeat units may outnumber the losses [39], that most of the mutations are single-step changes but changes involving two to five repeat units may also occur [39, 32], and that the mutation rate may be positively associated with the number of repeats [32]. As most of the observed changes in microsatellite length are by one or more repeat units the stepwise mutation model (SMM) has often been used to model microsatellite evolution [20, 11]. While the assumption may be that all microsatellite *loci* have the same evolutionary dynamics, it is now well understood that other mutational mechanisms may be present. Theoretical studies to describe genomic microsatellite distribution in which the models tested incorporated the competition between length dependent DNA replication slippage and base substitutions, interrupting the repeat, obtained better results than SMM alone [2, 21]. More recently Dieringer and Schlotterer [12] indicated that an additional mutational force, the length independent slippage, had to be incorporated to better describe genomic microsatellite distribution. Thus, differences among microsatellite alleles may involve several types of mutations.

In yeast, microsatellite *loci* have a considerable length variation and this polymorphism quickly made them attractive markers for a variety of analysis, including strain typing [15, 5, 25, 10, 4, 22, 37],

* Corresponding author: e-mail: cpais@bio.uminho.pt, Phone: +351 253604312

population structure [15, 16, 24, 18], and epidemiological studies [28, 34]. The dynamics of microsatellite mutations in yeast has been studied mainly in *Saccharomyces cerevisiae*. In these studies the authors analyzed the dependence of mutation rate on the length of the microsatellite and observed that the instability increased as repetitive tracts increased in size [41], and that interruptions in the repeat stabilized the microsatellites, lowering the mutation rates [31]. Additionally, they have analyzed several classes of trinucleotide repeats and observed a differential tendency to form secondary structures that escape DNA repair, thus presenting different instability [29].

In *Candida albicans* little is known about the contribution of the described mutational processes in the generation of allelic variability and sequence analysis of microsatellite allele in this species has been hardly documented [27]. Thus, the aim of this work was to determine the molecular nature of the length polymorphism observed at two compound microsatellites in *C. albicans* and infer the mechanisms that could rule that variability. Sequence analysis of the different alleles observed at two highly polymorphic microsatellites, CAI [34] and CEF3 [5] was performed in order to gain insights into the mechanisms underlying the intraspecific evolution of these microsatellites. A detailed study of the internal structure variation within the observed alleles was performed according to methodologies previously described [5, 6, 17, 19, 34]. The importance of these events regarding the application of these loci to strain differentiation in molecular epidemiology studies is also discussed.

2. CAI microsatellite analysis

2.1. Microsatellite sequencing and allele structure

PCR products obtained with CAI microsatellite consisted of fragments with different lengths, varying between 189 and 303 bp. Observed allele size variation was apparently due to either insertions or deletions of trinucleotide units. However, sequencing results revealed a much more complex scenario, as shown in Table 1. Several studies indicated that tracts of CAG can form secondary structures that increase microsatellite instability. This instability could also be observed in vivo, in yeast, where CNG triplets formed secondary structures that escaped DNA repair mechanisms [26, 29, 40]. In this view, the program *mfold* (version 3.1) [42] was used to analyse CAI complex alleles and identify patterns of variation. This analysis indicated that changes in microsatellite length could be due to the slippage of nonanucleotide??? repeats originated by formation of secondary structures stabilized by complementary base annealing that might escape DNA repair mechanisms (Figure 1 and Table 1).

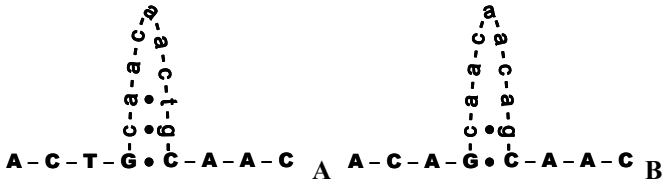

Fig. 1- Possible secondary structures with hairpin formation that could be formed in, (A) Variable Region 1 and (B) Variable Region 3, identified by the program mfold (version 3.1). Dots indicate the complementary base annealing that stabilizes the hairpins.

From this analysis six variable regions (VR) were identified (Table 1). The first variable region, (VR1), consisted of the presence of a nonanucleotide sequence (CAACAACTG) located upstream the repeated region that was present in all alleles except allele 11; the second, VR2, was the $(CAA)_n$ repeat motif present in all sequences, being the one that most contributed to allele size variation. VR3 was another nonanucleotide sequence (CAACAACAG) present in all alleles and was responsible for the more complex repeat structures. The region designated as VR4 was the less variable, all the alleles presented $(CAG)_2$ with just a few exceptions. Downstream the tandem repeat region two point mutations were identified at position 38, designated as VR5, and at position 108, as VR6.

Table 1 Sequence alignment of *C. albicans* CAI alleles, showing the defined variable regions VR1 to VR6. Group classification was based on variations at VR3 and Sub-Group on point mutations at VR5 and VR6.

Allele (bp): Sequence / VR1 /VR2/ VR3 /VR4/ VR5/ VR6/	Group	Sub-Group	Accession Number
11 (189):CTG $(CAA)_6$ (caacaacag) $(CAG)_2$----c------t	O	1	AY693670
11 (189):CTG $(CAA)_6$ (caacaacag) $(CAG)_2$----c------g	O	2	-
15 (201):CTG**caacaactg**$(CAA)_5$ (caacaacag) $(CAG)_4$----c------g	I	2	-
16 (204):CTGcaacaactg$(CAA)_5$ (caacaacag) $(CAG)_5$----t------t	I	3	AY693672
13 (195):CTGcaacaactg$(CAA)_2$ (**caacaacag**)$_2$ $(CAG)_2$----c-----g	II	2	AY693671
17 (207):CTGcaacaactg$(CAA)_6$ (caacaacag)$_2$ $(CAG)_2$----c-----g	II	2	AY693673
18 (210):CTGcaacaactg$(CAA)_7$ (caacaacag)$_2$ $(CAG)_2$----c-----t	II	1	AY693674
20a (216):CTGcaacaactg$(CAA)_9$ (caacaacag)$_2$ $(CAG)_2$----c-----t	II	1	AY693675
20 (216):CTGcaacaactg$(CAA)_9$ (caacaacag)$_2$ $(CAG)_2$----c-----g	II	2	-
21a (219):CTGcaacaactg$(CAA)_{10}$ (caacaacag)$_2$ $(CAG)_2$----c-----g	II	2	AY693676
21 (219):CTGcaacaactg$(CAA)_{10}$ (caacaacag)$_2$ $(CAG)_2$----c-----t	II	1	-
21b (219):CTGcaacaactg$(CAA)_{10}$ (caacaacag)$_2$ CA**A**CAG---c-----t	II	1	AY693677
22 (222):CTGcaacaactg$(CAA)_{11}$ (caacaacag)$_2$ $(CAG)_2$----c-----t	II	1	AY693678
23 (225):CTGcaacaactg$(CAA)_{12}$ (caacaacag)$_2$ $(CAG)_2$----c-----g	II	2	AY693679
24 (228):CTGcaacaactg$(CAA)_{13}$ (caacaacag)$_2$ $(CAG)_2$----c-----t	II	1	AY693680
25 (231):CTGcaacaactg$(CAA)_{14}$ (caacaacag)$_2$ $(CAG)_2$----c-----t	II	1	AY693681
26 (234):CTGcaacaactg$(CAA)_{15}$ (caacaacag)$_2$ $(CAG)_2$----c-----t	II	1	AY693682
27 (237):CTGcaacaactg$(CAA)_{16}$ (caacaacag)$_2$ $(CAG)_2$----c-----t	II	1	AY693683
28a (240):CTGcaacaactg$(CAA)_{17}$ (caacaacag)$_2$ $(CAG)_2$----c-----t	II	1	-
28 (240):CTGcaacaactg$(CAA)_{17}$ (caacaacag)$_2$ $(CAG)_2$----t-----t	II	3	AY693684
29a (243):CTGcaacaactg$(CAA)_{18}$ (caacaacag)$_2$ $(CAG)_2$----c-----t	II	1	-
29 (243):CTGcaacaactg$(CAA)_{18}$ (caacaacag)$_2$ $(CAG)_2$----t-----t	II	3	AY693685
30 (246):CTGcaacaactg$(CAA)_{19}$ (caacaacag)$_2$ $(CAG)_2$----c-----t	II	1	AY693686
32a (252):CTGcaacaactg$(CAA)_{21}$ (caacaacag)$_2$ $(CAG)_2$----c-----t	II	1	AY693687
33a (255):CTGcaacaactg$(CAA)_{22}$ (caacaacag)$_2$ $(CAG)_2$----c-----t	II	1	-
25 (231):CTGcaacaactg$(CAA)_{11}$ (**caacaacag**)$_3$ $(CAG)_2$----t-----t	III	3	-
26 (234):CTGcaacaactg$(CAA)_{12}$ (caacaacag)$_3$ $(CAG)_2$----t-----t	III	3	-
36 (264):CTGcaacaactg$(CAA)_{22}$ (caacaacag)$_3$ $(CAG)_2$----t-----t	III	3	AY693691
39 (273):CTGcaacaactg$(CAA)_{25}$ (caacaacag)$_3$ $(CAG)_2$----t-----t	III	3	AY693692
41 (279):CTGcaacaactg$(CAA)_{27}$ (caacaacag)$_3$ $(CAG)_2$----t-----t	III	3	-
42 (282):CTGcaacaactg$(CAA)_{28}$ (caacaacag)$_3$ $(CAG)_2$----t-----t	III	3	AY693693
45 (291):CTGcaacaactg$(CAA)_{31}$ (caacaacag)$_3$ $(CAG)_2$----t-----t	III	3	AY693694
46 (294):CTGcaacaactg$(CAA)_{32}$ (caacaacag)$_3$ $(CAG)_2$----t-----t	III	3	AY693695
47 (297):CTGcaacaactg$(CAA)_{33}$ (caacaacag)$_3$ $(CAG)_2$----t-----t	III	3	AY693696
49 (303):CTGcaacaactg$(CAA)_{35}$ (caacaacag)$_3$ $(CAG)_2$----t----t	III	3	AY693697
32b (252):CTGcaacaactg$(CAA)_{11}$(**caacaacag**)$_2$CAA(caacaacag)$_3$ $(CAG)_2$----t----t	IV	3	AY693688
33 (255):CTGcaacaactg$(CAA)_{12}$(caacaacag)$_2$CAA(caacaacag)$_3$ $(CAG)_2$----t----t	IV	3	AY693689
34 (258):CTGcaacaactg$(CAA)_{13}$(caacaacag)$_2$CAA(caacaacag)$_3$ $(CAG)_2$----t----t	IV	3	AY693690

Lowercase letters indicate the regions involved in secondary structures. Letters in bold indicate the nonanucleotide motif responsible for allele classification;- sequence not deposited in the GenBank

The analysis of the internal structure of *C. albicans* CAI alleles showed that this was a compound microsatellite with variable units of CAA and CAG. The length differences were mainly due to a variable number of CAA repeated units at VR2. Thus, regardless the differences in allele structure and the point mutations observed, as length differences were mainly due to trinucleotide variations, allele designation could be based on the number of trinucleotide repeats.

2.2. Mutation processes.

The allele variability found by sequence analysis of the CAI *locus* allowed the identification of various allelic groups and sub groups as a result of different mutational events (Table 1). Under a parsimony criterion, minimizing the overall number of mutational steps, it was possible to infer seven mutational steps explaining the observed allelic diversity (Figure 2).

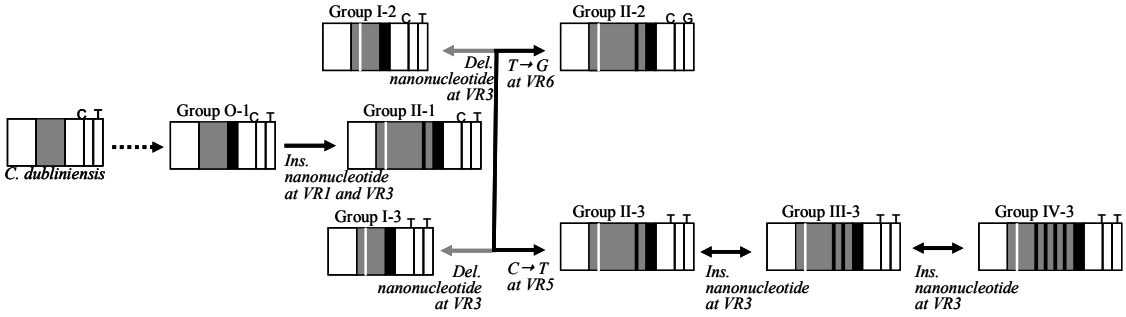

Fig. 2- Schematic hypothesis of mutational events observed in alleles from CAI microsatellite. From the hypothetical ancestral allele (Group O-1) insertion of the nonanucleotide at VR1 and VR3 by the formation of a loop generated a different allele structure (Group II-1). Two single-base substitutions seem to have occurred in this structure producing Groups II-2 and II-3. The formation of other loops in allele structure from Group II-3 created alleles observed in Group III-3 and Group IV-3. Occasionally, alleles from group II may lose the nonanucleotide at VR3 creating alleles from Group I-2 and I-3 (grey arrows).

Several studies supported the idea that interrupted alleles are derived from perfect ones [e.g. 7]. Alleles 11, 15 and 16, presented the simplest internal structure observed, and by microsatellite structure analysis alone could be considered hypothetical ancestral states. To address this question and since *C. dubliniensis* also presents this locus, DNA from *C. dubliniensis* strain ATCC MYA-646 was amplified with CAI primers and the corresponding fragment sequenced. Sequence alignment of *C. albicans* and *C. dubliniensis* alleles indicated that the original microsatellite was probably $(CAG)_n$ that degenerated into CAG/CAA in *C. albicans* (Figure 3).

```
C. albicans:     ATGCCATTGAGTGGAATTGGACTG..... (.........) .....CAGCAGCAGAATGCCCAACAGCAACTGCAAGTACCT
                 ||||| |||||||| |||||||           ||||  ||| | || ||||||||||  |||||||||
C. dubliniensis: ATGCCCTTGAGTGGTATTGGACCACAGCAGCAGCAGCCACAGCAAGCTCAACAGCAACAACAAGTACCT

C. albicans:     GTT------ATCCCAACACAAACACAAACATCACAACAAATGGCTTCAACTACCAATCACAAATCAGCT
                 |||      |  ||||||||||||||||||  |||||| |||||||||   |||||| ||||||||||
C. dubliniensis: GTTGTGGCCACACCAACACAAACACAAACAGGACAACAGATGGCTTCAGTAACCAATTACAAATCAGCT

C. albicans:     AAT---CTAATACCAGGGTTTTTACAAAACCCAACACAAGCCACT
                 |||   | ||||||||||||||| ||||||||||||||||||||
C. dubliniensis: AATCCTCCAATACCAGGGTTTTTCCAAAACCCAACACAAGCCACT
```

Fig 3- Sequence alignment of CAI locus from *C. albicans* and *C. dubliniensis*. Primer sequences are underlined, the microsatellite internal structure is presented in bold and the polymorphic positions VR5 and VR6 are highlighted.

The point mutations designated as VR5 and VR6 in *C. albicans,* in *C. dubliniensis* showed a C and a T, respectively. Therefore, taking into account the simplest structure and the point mutations at VR5 and VR6, the closest representative of *C. albicans* CAI original allele seems to be the one found in allele 11, (group O-1). If we take allele 16 or 15 as the representative of the ancestral state we would have to consider a deletion of VR1 followed by a T to C mutation at VR5 (allele 16) or a G to T mutation at VR6 (allele 15) to generate allele 11' pattern. On the other hand, the presence of a T in VR5 of more complex alleles would implicate a recurrent mutation, a phenomenon that is considered very rare.

Considering the two variable regions VR2 and VR3, alleles 13 and 17 to 32 presented the most frequently observed structure and harboured the highest level of variation at VR5 and VR6. This group of alleles (group II in Table 1) was probably derived from allele 11' structure by first, an expansion of the nonanucleotide (CAACAACTG) at VR1 and then, an expansion of the nonanucleotide (CAACAACAG) at VR3, generating the structure represented in Table 1 as group II-1 (Figure 2). In this sequence background two different mutations seem to have occurred downstream the repetitive sequence, a C to T mutation (group II-3) and a T to G mutation (group II-2), accounting for the variation found at jVR5 and VR6, respectively. Allelic variability within this group was further created by expansion of the trinucleotide repeat $(CAA)_n$ at VR2.

Alleles from group III originated from Group II by expansion of other nonanucleotide (CAACAACAG) at VR3, and alleles from group IV derived from Group III by formation of another loop at VR3. All alleles from groups III and IV presented the haplotype TT at VR5 and VR6, indicating that they derived from the structure of group II-3 alleles. Occasionally, alleles from group II may lose the nonanucleotide at VR3, creating alleles from Group I-2 and I-3. Alleles from Group I-1 were not detected in this study.

Analysis of CAI allelic frequencies showed that alleles belonging to group II were the most frequent (Figure 4A). This observation supported the hypothesis that the major mutational process driving allelic variability at this microsatellite was the addition/deletion of CAA units at VR2. Curiously, although alleles from group III presented the longest simple tracts at VR2, their frequencies were much lower than would be expected from a positive correlation between mutational rate and repeat number. This result seems to indicate that the presence of CAG units in the longer tracts of CAA greatly stabilized the mutation rate at VR2.

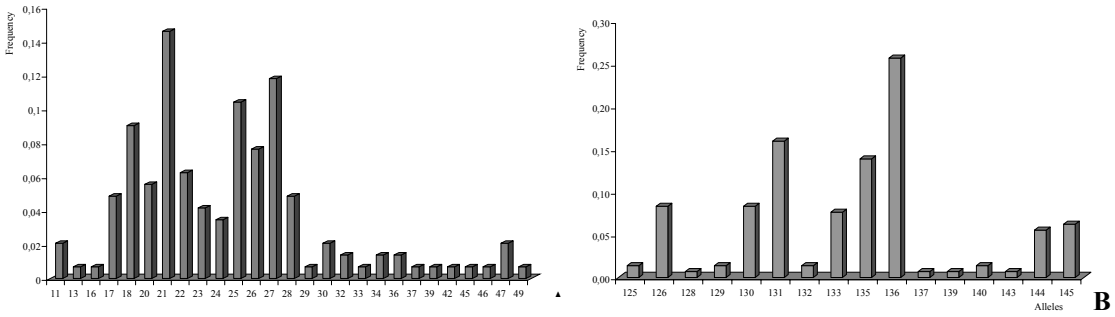

Fig. 4- CAI (A) and CEF3 (B) allele frequencies.

In summary, for CAI microsatellite analysis we have detected three levels of mutational events: (i) addition/deletion of CAA repeats in the variable region VR2, (ii) addition/deletion of nanonucleotide units with the formation of secondary structures at VR3, and (iii) point mutations at positions 38 (VR5) and 108 (VR6) downstream the repeated region.

3. CEF3 microsatellite analysis

3.1. Sequencing and allele structure

A total of 16 different alleles were found at CEF3 locus varying in size from 125 to 145 bp. Sequencing results of 14 of the alleles are shown in Table 2. The consensus structure obtained for CEF3 microsatellite locus was in accordance with accession number Z11484 deposited in the NCBI database. Alleles observed varied by one up to four bp, which is consistent with simultaneous variations at the tri and tetra units.

A detailed analysis of the flanking sequences of CEF3 locus revealed a 13 bp variable tract, 10 bp upstream the tandem repeat units, including the regions designated as VR1 and VR2 (Table 2). The tri and tetra repeats were designated as VR4 and VR3, respectively.

Table 2. Sequencing data and alignment of *C. albicans* CEF3 alleles as well as variable regions VR1 to VR4. Z11484 is the consensus sequence.

Allele: Variable Regions /VR1-------VR2-/----------/-VR3-/-VR4 /	**Group**	**Sub-group**	**Accession number**
Z11484(aaatttttttttcc)tttcttataa(TTTC)$_5$(TTC)$_5$ 34bp			
126 (**a**---------ttc-)----------(TTTC)$_4$(TTC)$_3$	I	1	AY693698
131 (a---------ttc-)----------(TTTC)$_6$(TTC)$_2$	I	1	AY693702
130 (a---------tt**t**-)----------(TTTC)$_5$(TTC)$_3$	II	1	AY693701
128 (---------ttc-)----------(TTTC)$_4$(TTC)$_4$	III	1	AY693699
137 (---------ttc-)----------(TTTC)$_4$(TTC)$_2$(TT**T**)(TTC)$_4$	III	2	AY693706
129 (---------ttc-)----------(TTTC)$_5$(TTC)$_3$	III	1	AY693700
145 (---------ttc-)----------(TTTC)$_6$(TTC)$_7$	III	1	AY693711
135 (---------ttc-)----------(TTTC)$_5$(TTC)$_5$	III	1	AY693704
136 (---------ttc-)----------(TTTC)$_6$(TTC)$_4$	III	1	AY693705
140 (---------ttc-)----------(TTTC)$_7$(TTC)$_4$	III	1	AY693708
143 (---------ttc-)----------(TTTC)$_7$(TTC)$_5$	III	1	AY693709
144 (---------ttc-)----------(TTTC)$_8$(TTC)$_4$	III	1	AY693710
133 (a---------t c-)----------(TTTC)$_6$(TTC)$_3$	IV	1	AY693703
139 (a--------- c-)----------(TTTC)$_7$(TTC)$_4$	V	1	AY693707

Lowercase letters indicate the flanking regions and capital letters the microsatellite region. Bold letters indicate modifications outside the microsatellite regions related to the process of allele evolution. The symbol – indicates the same nucleotide and the absence indicates deletion.

3.2. Mutation processes

Analysis of the variability observed at the CEF3 locus allowed the inference, under the parsimony criterion, of five mutational steps to explain this observed allelic diversity (Figure 5). In order to determine the hypothetical ancestral state of these CEF3 alleles, only the point mutations at VR1 and VR2 were considered. The structure observed in alleles 126 and 131 (Group I in Table 2) was regarded as the hypothetical ancestral since it presented the highest variability associated. Furthermore, considering this structure as the hypothetical ancestral, invoking recurrent mutations was avoided. In the background of this sequence structure, three different mutations seem to have occurred that generated the majority of allelic structures. A C to T mutation appeared in allele 130 (Group II), a deletion of a T in allele 133 (Group IV) and a deletion of an A was observed in the remaining alleles (Group III). The VR2 structure observed in allele 139 seems to have derived from allele 133 structure by the deletion of another T (Group V). No differences in repeat structure were observed at VR3 and VR4, except for allele 137 where a C to T mutation in the trinucleotide block was found (Group III-2). For the majority of the alleles the structure also differed in the number of tetranucleotide and/or trinucleotide motifs, except in the cases of alleles 139 and 140, and alleles 129 and 130, in which the size differences were due to variations not at the microsatellite but to indels at the upstream flanking regions, VR1 and VR2 (Table 2).

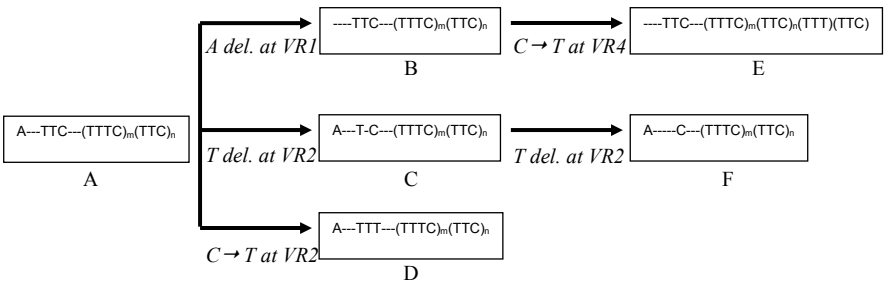

Fig. 5- Schematic hypothesis of mutational events observed in alleles from CEF3 microsatellite. From the hypothetical ancestral allele, group I (A) three single-base mutations seem to have occurred originating different allele structures, group III (B), group IV (C) and group II (D). Allele structures present in group III-2 (E) and V (F) derived by single base mutations from groups III and IV, respectively.

The comparison of CEF3 alleles observed in this study with the ones previously described was only based in VR3 and VR4, since no reference to other variable regions was reported. Considering the molecular length ten alleles were similar to the ones reported in the literature, but looking at their repetitive structure only seven could be considered as common [5, 10]. Alleles 131, 137 and 139 although having the same molecular length as alleles previously described presented different sequence structure at VR3 and VR4.

Allelic diversity at CEF3 appears to be the result of the combination of only two distinct mutational events (i) changes in the number of microsatellite units at VR3 and VR4, and (ii) indels outside the repeated region, at VR1 and VR2. Analysis of CEF3 allele frequencies does not indicate particular groups of alleles as more frequent, suggesting that the two distinct mutational events identified may have a similar contribution to allelic variability in this microsatellite (Figure 4B).

4. Discussion

Since the emergence of microsatellite analysis as a PCR based method they have been successfully used as tools for an increasingly diverse range of studies. Despite the remarkable amount of works the processes by which microsatellite mutate, and hence evolve, are not fully understood. Among several mutational processes, the stepwise mutation model (SMM) seems to be the one responsible for the major length polymorphism observed in this kind of markers. Since several studies indicate important departures from this strict model, the question of whether SMM would be the only mechanism leading to length changes gain interest. Sequence studies of microsatellite variability, both within and between species, confirmed that changes in allele length are most often due to size alterations in the repeated region, but the introduction of interruptions in the repeated sequence and homoplasy are quite common [14, 1]). Sequence studies also revealed mutations in regions flanking the microsatellite [1].

In *C. albicans* microsatellite variability has been successfully used as a tool to differentiate strains in molecular epidemiology and population studies. However, patterns of intraspecific microsatellite evolution have not been widely addressed. In this study we have looked at allele sequence and inferred the mutational processes underlying the observed polymorphism at CAI and CEF3 microsatellites. Sequence analysis revealed high allelic variability within the microsatellite and flanking regions of both CAI and CEF3 *loci* that involved multiple mutational processes, which cannot be explained just by the SMM. According to this model only the addition/deletion of the repeated units is responsible for length variations, however several authors indicate that differences among alleles may involve additional types of mutations [1, 30, 8, 12]. For CAI microsatellite, the overall allelic diversity appears to be the result of the combination of three distinct types of mutational events: (i) addition/deletion of CAA repeats in the variable region VR2, (ii) addition/deletion of nanonucleotide units with the formation of secondary structures at VR3, and (iii) point mutations at positions 38 (VR5) and 108 (VR6) downstream the

repeated region. The most frequent alleles differed in the number of trinucleotide repeats indicating that the predominant mutational event in CAI was the addition/deletion of CAA units at VR2. These observations are in agreement with other studies in which different mutational processes were observed but the predominant within a species were the indels of the smaller repeated units [14, 1, 31]. The presence of interruptions, in VR3, seems to have as a consequence the stopping of the indefinite expansion of the microsatellite and it can explain the more complex structural patterns present in the larger alleles. This was demonstrated in yeast by the study of Petes *et al.* [31] in which the incorporation of a single variant within a 51 bp poly GT tract stabilized the microsatellite about five fold. The observed allele frequencies for CAI are consistent with this hypothesis, since alleles with 3 to 5 CAG interruption presented a much lower frequency. Bichara *et al*. [31] suggested that mutational intermediates from different regions of an interrupted repeat isomerise to form large deletions making the allele almost unrecognizable. Our observations agree not only with the microsatellite stabilization but also with the hypothesis that interruptions can lead to repeat shortening [39, 31, 38]. Alleles 15 and 16 presented a microsatellite structure that may be the result of this type of shortening events.

CEF3 microsatellite analysis revealed two different mutational events that contributed to allele variability, (i) addition/deletion of TTTC and TTC repeats in the variable region VR3 and VR4; and (ii) variability at a 13 bp tract, 10 bp upstream the tandem repeated region (VR1 and VR2). The first level of polymorphism had been initially described by Bretagne *et al*. [5] and reported as the responsible for length variation among CEF3 alleles. This variation could be due to differences in the number of TTTC as well as in the TTC units, and alleles with the same size could present different allelic internal structures. The second level of polymorphism, described in the present study, was found to be due to four point mutation events giving rise to five different allele structures. This region presented nucleotide insertions/deletions generating length variability independently from repeat variation.

The present study provides evidence for the mutational events that generated allele length variability at CAI and CEF3 microsatellite *loci* in *C. albicans*, indicating that the detectable patterns of microsatellite evolution may differ between *loci*. Although mutational events driving CEF3 evolution were similar to the ones observed for CAI, analysis of CEF3 allele frequencies seems to indicate that the two different mutational events may have an equal contribution to variation in CEF3 alleles.

A factor limiting the usefulness of microsatellites for genetic analysis, particularly strain identity using standard genotyping procedures, is the possible convergence of ancestral alleles to the same length by different mutational events, an effect known as homoplasy [14, 1, 27]. In CAI we have identified homoplasy due to variation in different parts of the compound microsatellite, for e.g. alleles 132a and 132b and alleles 121a and 121b. In the microsatellite CEF3 no homoplasy was detected in the alleles found in this study. However, when these alleles were compared with alleles sequenced in other studies [5, 10] homoplasy was observed between them and was also due to variations in different parts of the compound microsatellite. As supported by the study of Estoup *et al.* [14], homoplasy does not represent a significant problem for strain differentiation but for strain relationship may represent a problem. However, when appropriate population models, as well as several loci are considered, the high microsatellite variability often largely compensates for their eventual homoplasic evolution. Thus, the application of microsatellite loci in studies such as molecular epidemiology and population studies is highly recommended since the accurate discrimination of genetically divergent groups within pathogenic species is critical to the appropriate development and use of treatment strategies.

Acknowledgements This study was supported by a grant from the Fundação para a Ciência e Tecnologia (POCI/SAU-IM/58014/2004). We thank Magda Graça for technical assistance with the sequence analysis. We also gratefully acknowledge all the colleagues who have supplied the clinical isolates used in this study. IPATIMUP is partially funded by the VI Framework Program "Programa Operacional Ciência e Inovação (POCI 2010).

References

[1] ANGERS, B. and L. BERNATCHEZ, 1997 Complex evolution of a salmonid microsatellite locus and its consequences in inferring allelic divergence from size information. Mol Biol Evol. 14: 230-8.
[2] BELL, G. I., and J. JURKA, 1997. The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single-step mutation process. J. Mol. Evol. 44: 414-421.
[3] BICHARA, M., S. SCHUMACHER and R. P. FUCHS, 1995 Genetic instability within monotonous runs of CpG sequences in Escherichia coli. Genetics. 140: 897-907.
[4] BOTTEREL, F., C. DESTERKE, C. COSTA and S. BRETAGNE, 2001 Analysis of microsatellite markers of Candida albicans used for rapid typing. J Clin Microbiol. 39: 4076-81.
[5] BRETAGNE, S., J. M. COSTA, C. BESMOND, R. C. CARSIQUE and R. CALDERONE, 1997 Microsatellite polymorphism in the promoter sequence of the elongation factor 3 gene of Candida albicans as the basis for a typing system. J Clin Microbiol 35: 1777-1780.
[6] BUDOWLE, B., R. CHAKRABORTY, A. M. GIUSTI, A. J. EISENBERG and R. C. ALLEN, 1991 Analysis of the VNRT locus D1S80 by the PCR followed by high resolution PAGE. Am. J. Hum. Genet. 48: 137-144.
[7] BULL, L. N., C. R. PABON-PENA and N. B. FREIMER, 1999 Compound microsatellite repeats: practical and theoretical features. Genome Res. 9: 830-838.
[8] COLSON, I. and D. B. GOLDSTEIN, 1999 Evidence for complex mutations at microsatellite loci in Drosophila. Genetics. 152: 617-27.
[9] CORREIA, A., P. SAMPAIO, J. ALMEIDA and C. PAIS, 2004 Identification of Candida clinical isolates by PCR fingerprinting: a contribution to the study of molecular epidemiology of candidiasis in Portugal. J Clin Microbiol 42:5899-5903.
[10] DALLE, F., N. FRANCO, J. N. LOPEZ, O. J. VAGNER, D. CAILLOT, P. CHAVANET, B. CUISENIER, S. AHO, S. LIZARD and A. BONNIN, 2000 Comparative genotyping of Candida albicans bloodstream and nonbloodstream isolates at a polymorphic microsatellite locus. J Clin Microbiol. 38: 4554-9.
[11] DI RIENZO, A., A. C. PETERSON, C. GARZA, A. M. VALDES, M. SLATKIN and N. B. FREIMER, 1994 Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. 91: 3166-3170.
[12] DIERINGER, D. and C. SCHLOTTERER, 2003 Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res. 13: 2242-51.
[13] ELLEGREN, H. 2000 Microsatellite mutations in the germline: implications for evolutionary inference. Trends Genet. 16: 551-8.
[14] ESTOUP, A., C. TAILLIEZ, J. M. CORNUET and M. SOLIGNAC, 1995 Size homoplasy and mutational processes of interrupted microsatellites in two bee species, Apis mellifera and Bombus terrestris (Apidae). Mol Biol Evol. 12: 1074-84.
[15] FIELD, D., L. EGGERT, L. METZGAR, R. ROSE and C. WILLS, 1996 Use of polymorphic short and clustered coding-region microsatellites to distinguish strains of Candida albicans. FEMS Immunol Med Microbiol. 15: 73-9.
[16] FUNDYGA, R. E., T. J. LOTT and J. ARNOLD, 2002 Population structure of Candida albicans, a member of the human flora, as determined by microsatellite loci. Infect Genet Evol. 2: 57-68.
[17] GUSMÃO L., M. J. PRATA, A. AMORIM, F. SILVA and I. BESSA, 1997. Characterization of four short tandem repeat loci (TH01, VWA31/A, CD4 and TP53) in Northern Portugal. Hum. Biol. 69: 31-40.
[18] HOWELL, K. S., E. J. BARTOWSKY, G. H. FLEET, and P. A. HENSCHKE, 2004 Microsatellite PCR profiling of Saccharomyces cerevisiae strains during wine fermentation. Lett. Appl. Microbiol.38: 315-320.
[19] KAISER, C., S. MICHAELIS and A. MITCHELL, 1994. Methods in yeast genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, N. Y.
[20] KIMURA, M. and T. OHTA, 1973 Eukaryotes-prokaryotes divergence estimated by 5S ribosomal RNA sequences. Nat New Biol. 243: 199-200.
[21] KRUGLYAK, S., R. DURRETT, M. D. SCHUG, and C. F. AQUADRO, 2000 Distribution and abundance of microsatellites in the yeast genome can Be explained by a balance between slippage events and point mutations. Mol. Biol. Evol. 17: 1210-1219.
[22] LASKER, B. A., and Y. RAN, 2004 Analysis of polymorphic microsatellite markers for typing Penicillium marneffei isolates. J. Clin. Microbiol. 42: 1483-1490.
[23] LEVINSON, G. and G. A. GUTMAN, 1987 Slipped-strand misrepairing: A major mechanism for DNA sequence evolution. Mol. Biol. Evol. 4: 203-221.

[24] LOTT, T. J., R. E. FUNDYGA, M. E. BRANDT, L. H. HARRISON, A. N. SOFAIR et al., 2003 Stability of allelic frequencies and distributions of Candida albicans microsatellite loci from U.S. population-based surveillance isolates. J Clin Microbiol. 41: 1316-21.
[25] LUNEL, F. V., L. LICCIARDELLO, S. STEFANI, H. A. VERBRUGH, W. J. MELCHERS, et al., 1998 Lack of consistent short sequence repeat polymorphisms in genetically homologous colonizing and invasive Candida albicans strains. J Bacteriol. 180: 3771-8.
[26] MAURER DJ, BL O'CALLAGHAN, and DM LIVINGSTON. 1998. Mapping the polarity of changes that occur in interrupted CAG repeat tracts in yeast. Mol Cell Biol. 18:4597-4604.
[27] METZGAR, D., D. FIELD, R. HAUBRICH, and C. WILLS, 1998 Sequence analysis of a compound coding-region microsatellite in Candida albicans resolves homoplasies and provides a high-resolution tool for genotyping. FEMS Immunol. Med. Microbiol. 20: 103-109.
[28] MEYER, W., K. MARSZEWSKA, M. AMIRMOSTOFIAN, R. P. IGREJA, C. HARDTKE et al., 1999 Molecular typing of global isolates of Cryptococcus neoformans var. neoformans by polymerase chain reaction fingerprinting and randomly amplified polymorphic DNA-a pilot study to standardize techniques on which to base a detailed epidemiological survey. Electrophoresis. 20: 1790-1799.
[29] MOORE, H., P. W. GREENWELL, C. P. LIU, N. ARNHEIM, and T. D. PETES. 1999. Triplet repeats form secondary structures that escape DNA repair in yeast. Proc. Natl. Acad. Sci. 96: 1504-1509.
[30] ORTI, G, D., E. PEARSE and J. C. AVISE, 1997 Phylogenetic assessment of length variation at a microsatellite locus. Proc Natl Acad Sci. 94: 10745-9.
[31] PETES, T. D., P. W. GREENWELL and M. DOMINSKA, 1997 Stabilization of microsatellite sequences by variant repeats in the yeast Saccharomyces cerevisiae. Genetics. 146: 491-8.
[32] PRIMMER, C. R., N. SAINO, A. P. MØLLER, and H. ELLEGREN, 1996 Directional evolution in germline microsatellite mutations. Nat. Genet. 13: 391-393.
[33] PUPKO, T. and D. GRAUR, 1999 Evolution of microsatellites in the yeast Saccharomyces cerevisiae: role of length and number of repeated units. J Mol Evol. 48: 313-6.
[34] SAMPAIO, P., L. GUSMÃO, C. ALVES, C. PINA-VAZ, A. AMORIM and C. PAIS, 2003 Highly polymorphic microsatellite for identification of Candida albicans strains. J Clin Microbiol. 41: 552-7.
[35] SCHLOTTERER, C. and D. TAUTZ, 1992. Slippage synthesis of simple sequence DNA. Nuc. Acid Res. 20: 211-215.
[36] SCHLOTTERER, C., 2000 Microsatellite analysis indicates genetic differentiation of the neo-sex chromosomes in Drosophila americana. Heredity. 85: 610-6.
[37] SHEMER, R., Z. WEISSMAN, N. HASHMAN, and D. KORNITZER, 2001 A highly polymorphic degenerate microsatellite for molecular strain typing of Candida krusei. Microbiology. 147: 2021-2028.
[38] TAYLOR, J. S., J. M. DURKIN and F. BREDEN, 1999 The death of a microsatellite: a phylogenetic perspective on microsatellite interruptions. Mol Biol Evol. 16: 567-72.
[39] WEBER, J. L. and C. WONG, 1993 Mutation of human short tandem repeats. Hum. Mol. Genet. 2: 1123-1128.
[40] WELLS R.D. 1996. Molecular basis of genetic instability of triplet repeats. J. Biol. Chem. 271:2875-2878. Review
[41] WIERDL, M., M. DOMINSKA, and T. D. PETES, 1997 Microsatellite instability in yeast: dependence on the length of the microsatellite. Genetics. 146: 769-779.
[42] ZUKER, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31: 3406-3415