Phylogeny and evolution of the aspartyl protease family from clinically relevant *Candida* species

B Parra-Ortega, H Cruz-Torres, L Villa-Tanaca, C Hernández-Rodríguez/+

Departamento de Microbiología, Escuela Nacional de Ciencias Biológicas, Instituto Politécnico Nacional, Plan de Ayala y Prol. Carpio, Colonia Casco de Santo Tomás, CP 11340 Agencia de Correos 220, México, DF, México

Aspartyl proteases are a class of enzymes that include the yeast aspartyl proteases and secreted aspartyl protease (Sap) superfamilies. Several Sap superfamily members have been demonstrated or suggested as virulence factors in opportunistic pathogens of the genus Candida. Candida albicans, Candida tropicalis, Candida dubliniensis *and* Candida parapsilosis *harbour 10, four, eight and three* SAP *genes, respectively. In this work, genome mining and phylogenetic analyses revealed the presence of new members of the Sap superfamily in* C. tropicalis *(8),* Candida guilliermondii *(8),* C. parapsilosis *(11) and* Candida lusitaniae *(3). A total of 12 Sap families, containing proteins with at least 50% similarity, were discovered in opportunistic, pathogenic* Candida *spp. In several Sap families, at least two subfamilies or orthologous groups were identified, each defined by > 90% sequence similitude, functional similarity and synteny among its members. No new members of previously described Sap families were found in a* Candida *spp. clinical strain collection; however, the universality of* SAPT *gene distribution among* C. tropicalis *strains was demonstrated. In addition, several features of opportunistic pathogenic* Candida *species, such as gene duplications and inversions, similitude, synteny, putative transcription factor binding sites and genome traits of* SAP *gene superfamily were described in a molecular evolutionary context.*

Key words: pathogenic *Candida* spp. - aspartyl proteases phylogeny - evolution

The polyphyletic genus *Candida* includes fungal species pathogenic in humans, including *Candida albicans* (Cal), *Candida dubliniensis* (Cdu), *Candida glabrata* (Cgl), *Candida guilliermondii* (Cgu)*, Candida lusitaniae* (Clu), *Candida parapsilosis* (Cpa) and *Candida tropicalis* (Ctr). These species are together responsible for 34% of opportunistic fungal infections (Pfaller & Diekema 2007) and *C. albicans* is the most common fungal pathogen in humans (Moran et al. 2002). Several *Candida* virulence factors have been discovered or proposed; these include adhesion, hyphal formation, phenotypic switching and production of extracellular hydrolytic enzymes such as phospholipases, lipases and secreted aspartyl proteinases (Saps) (Naglik et al. 2003).

Aspartyl proteases are considered among the most important virulence factors in *C. albicans*. The Sap pre-pro-enzymes contain an amino-terminal signal peptide, two conserved active site aspartate residues and four cysteine residues implicated in the maintenance of the three-dimensional structure (Fig. 1). The mature enzyme is obtained after a successive proteolytic processing performed by a signal peptidase and the Kex2 enzyme (Hube & Naglik 2002). Ten aspartyl proteases have been identified in *C. albicans* (Sap1-Sap10), four in *C. tropicalis* (Sapt1-Sapt4), eight in *C. dubliniensis* (Sapcd1-Sapcd4; Sapcd7-Sapcd10) and three in *C. parapsilosis* (Sapp1-Sapp3) (Monod et al. 1994, Pichova et al. 2001, Zaugg et al. 2001). Sap9 and Sap10 of *C. albicans* have C-terminal consensus sequences for glycosylphosphatidylinositol (GPI) modification, similar to yapsin proteins (Monod et al. 1998). No typical Saps have been described in *C. glabrata,* a species more phylogenetically related to *Saccharomyces cerevisiae* than other pathogenic *Candida* species. However, both *C. glabrata* and *S. cerevisiae* possess several yapsins, which are non-secreted, GPI-anchored aspartyl proteases involved in cell wall integrity and cell-cell interactions (Krysan et al. 2005).

Previous studies have described a phylogenetic relationship between Saps from *C. albicans*, *C. tropicalis* and *C. parapsilosis,* including sequences available at that time. These proteins were classified into three groups based on the sequences available at the time: Sap1-3, Sap4-6 and Sap7, Sap9 and Sap10 (Sther et al. 2000). However, only Sap1-3 and Sap4-6 were closely clustered; Sap8 was remotely related to both clusters, Sap9 and Sap10 were clustered between them and Sap7 was externally positioned. Phylogenetic trees of *C. tropicalis*, *C. parapsilosis* and *C. albicans* Saps were also constructed and the clustering topology of Saps from *C. albicans* was essentially maintained. *C. tropicalis* Sapt4 was externally clustered with *C. albicans* Sap1-Sap3; *C. tropicalis* Sapt1 clustered with *C. albicans* Sap8, *C. tropicalis* Sapt2 and Sapt3 formed independent branches and *C. parapsilosis* Sapp1 and Sapp2 grouped together, distant from the other groups (Monod et al. 1998, Hoyer et al. 2001, Zaug et al. 2001).

Financial support: CONACyT-14299, CONACyT-69984, SIP-IPN 2008-2350, SIP-IPN 2007-0248. BPO and HCT are fellows from CONACyT, BPO is PIFI-IPN fellow, LVT and CHHR received support from COFAA-IPN and EDI-IPN.
+ Corresponding author: chdez38@hotmail.com
Received 15 July 2008
Accepted 23 March 2009

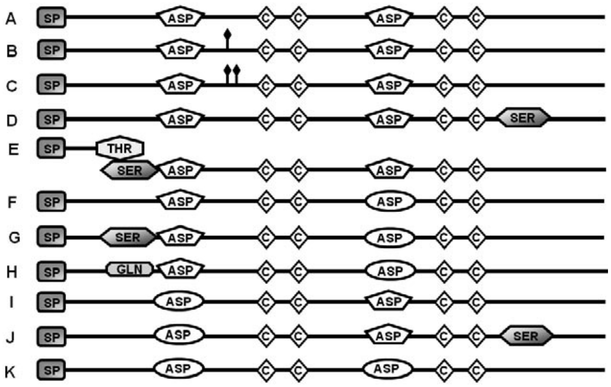

Fig. 1: motifs of *Candida* spp. secreted aspartyl proteinases. Rectangle boxes (SP): amine terminal signal peptide; pentagon (ASP): aspartyl protease domains in agreement with PROSITE ([LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-[STAPDENQ]-{GQ}-[LIVMFSTNC]-{EGK}-[LIVMFGTA]) and Softberry ([LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-[STAPDENQ]-x-[LIVMFSTNC]-x-[LIVMFGTA]) motifs; circles (ASP): atypical aspartyl protease domains; hexagons: serine (SER), threonine (THR) or glutamine (QLN) rich regions; arrows: RGD motif; rhombus: cysteine (C) residues. Cal: *Candida albicans*; Cdu: *Candida dubliniensis*; Cgl: *Candida glabrata*; Cgu: *Candida guilliermondii*; Clu: *Candida lusitaniae*; Cpa: *Candida parapsilosis*; Ctr: *Candida tropicalis*. Cal Sap1 (CaO19_5714), Cal Sap2 (CaO19.3708), Cal Sap3 (CaO19_6001), Cdu Sapd1 (Cd36_63430), Cdu Sapd2 (Cd36_33460), Cdu Sapd3 (Cd36_85220), Cdu Sapd4 (Cd36_63420), Cdu Sapd9 (Cd36_83850), Ctr Sapt1 (X61438), Ctr Sapt2 (AF115320), Ctr Sapt3 (AF115321), Ctr Sapt4 (AF115322), Ctr (CTRG_02711), Ctr (CTRG_02698), Cpa Sapp1 (Z11919), Cpa Sapp2 (Z11918), Cpa Sapp3 (AF339513), Cpa (CPAG_04800), Cpa (CPAG_03615), Cpa (CPAG_01799), Cpa (CPAG_03663); Cgu (PGUG_03957), Cgu (PGUG_03959); Cgu (PGUG_02701), Cgu (PGUG_00002), Scer Yap1 (YLR120C); B) Cal Sap4 (CaO19_5716), Cal Sap5 (CaO19_13032); C) Cal Sap6 (CaO19_12988); D) Ctr (CTRG_02621), Ctr (CTRG_03508), Cpa (CPAG_01798), Cpa (CPAG_01800), Cpa (CPAG_02614), Cgu (PGUG_01925), Clu (CLUG_03179); E) Ctr (CTRG_04491); F) Cdu Sapd7 (Cd36_04610), Cdu Sapd8 (Cd36_82480), Ctr (CTRG_02710), Cpa (CPAG_00343), Cpa (CPAG_02617), Cgu (PGUG_02700), Clu (CLUG_05711); G) Cal Sap7 (CaO19.8376), Clu (CLUG_04798); H) Cal Sap8 (CaO19.7872); I) Cgu (PGUG_03958), Cal Sap9 (Ca19.14190), Cal Sap10 (Ca19.11320), Ctr (CTRG_00264), Cpa (CPAG_05341), Cpa (CPAG_04721), Cpa (CPAG_04722); J) Cdu Sapd10 (Cd36_44180); K) Ctr (CTRG_00096), Cgu (PGUG_02218). Proteins and motifs are not drawn to scale.

Recently, the genomes of the main pathogenic *Candida* species were sequenced and released, although full gene annotation is not yet complete. In this work, we attempted to define orthologous and paralogous groups of Saps from pathogenic *Candida* species. We sought to study the distribution and abundance of these genes among species and to define evolutionary traits using a phylogenetic approach. We also aimed to define superfamilies, families and subfamilies of *Candida* Saps.

MATERIALS AND METHODS

Identification of Candida SAP genes in silico - Saps consist of two domains. Each domain contains an active site centred on a catalytic aspartyl residue and four conserved cysteine residues responsible for two conserved disulfide bridges (Hube & Naglik 2002). Sequences corresponding to aspartyl proteases from *S. cerevisiae* and *C. albicans* were obtained from *Saccharomyces* (http://www.yeastgenome.org) and *Candida* Genome Databases (http://www.candidagenome.org), respectively. *C. dubliniensis* sequences were from the Sanger Institute Microorganisms Sequencing Group (http://www.sanger.ac.uk/sequencing/Candida/dubliniensis/). Sequences from *C. guilliermondii*, *C. lusitaniae*, *C. tropicalis* and *C. parapsilosis* were obtained from the Candida Database (http://www.broad.mit.edu/annotation/genome/candida_group/MultiHome.html). The GenBank database (http://www.ncbi.nlm.nih.gov) was also used. Yapsins, GPI-linked aspartyl proteases considered a separate superfamily, were excluded from this analysis (Krysan et al. 2005).

Bioinformatics analysis - Basic sequence manipulations were performed with DNAMAN version 3.0 (Lynnon BioSoft 1994-1997). Prediction of putative transcription factor binding sites was performed using MatInspector version 2.2 (www.genomatix.de/products/MatInspector/) (Quandt et al. 1995). Theoretical isoelectric points and molecular weights were calculated using Antheprot 2000 version 5.2. Prediction of motif sequences was performed with PROSITE (http://www.expasy.org) (Falquet et al. 2002). PSORTII (http://www.psort.org/) and Softberry (http://www.softberry.com) were used to predict subcellular localization; Softberry was also used to find exons. Protein internal sequence repeats in proteins were detected by the TRUST repeat detection method (Szklarczyk & Heringa 2004).

Strains and culture media - The *Candida* spp. strains used in this work were obtained from Departamento de Microbiología, ENCB Collection. In addition, *C. albicans* ATCC 10231 and *C. dubliniensis* CD36 and CD92 strains were donated by Dr. Derek J. Sullivan (University of Dublin) and *C. glabrata* CBS138 and BG6 strains were donated by Dr. Bernard Dujon (Institut Pasteur and Université Pierre et Marie Curie). Aspartyl protease gene diversity was studied in various clinical strains of *C. albicans* (3 strains), *C. glabrata* (3), *C. tropicalis* (90), *C. lusitaniae* (3), *C. guilliermondii* (3), *C. parapsilosis* (8), *Candida krusei* (3), *Candida kefyr* (3) and *S. cerevisiae* (1). The yeast were routinely grown on Sabouraud dextrose agar plates at 37°C for 24 h and were stored on YPD plates (1% yeast extract, 2% peptone, 2% glucose and 2% agar) at 4ºC.

DNA extractions - Yeast cultures were grown in YPD broth overnight at 37°C with shaking at 150 rpm. DNA was extracted from washed yeast cultures using a previously reported protocol (Hoffman & Winston 1987). DNA concentrations and *A*260/*A*280 ratios were determined with a UV-visible spectrophotometer (Varian Cary 50, Australia). A DNA sample with an *A*260/*A*280 ratio of 1.8-2.0 was considered acceptable for further use.

PCR procedures - Primers were designed to amplify *SAP* genes by from four families: family I (*SAP1*-*SAP3* from *C. albicans* and *SAPT4* from *C. tropicalis*), family II (*SAP4*-*SAP6* from *C. albicans*), family III (*SAP8* from

C. albicans and *SAPT1* from *C. tropicalis*) and family IV (*SAPP1-SAPP3* from *C. parapsilosis*). Primer pairs used were: family I, sense (S): 5′ -TACTGGATCWTCTGATYTATGGG-3′ and antisense (AS): 5′ -GCTGATCTYAAAAAGTTATGACCAAGAAT-3′; family II, S: 5′ -GTTTGTTACCTTAGACTTTAATGTC-3′ and AS: 5′ -TTGTCACCAAGAATATTATCTTCAC-3′; family III, S: 5′ -TTGCTTTGTTTGCHCAAGGT-3′ and AS: 5′ -AGTGTAYTTGACTTGAGCCA-3′; and family IV, S: 5′ -TTGCTTTGTTTGCHCAAGGT-3′ and AS: 5′ -TTKGCATTATCAACACCACC-3′. All PCRs were performed in a buffer containing 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 2 mM $MgCl_2$, each deoxynucleoside triphosphate at a concentration of 0.2 mM, each primer at a concentration of 0.6 μM, 4 ng/μL of genomic DNA, and 1.5 U/μL of *Taq* polymerase (Invitrogen). The PCR conditions included a denaturation step for 3 min at 94°C, followed by 38 amplification cycles consisting of 1 min at 94°C, 1 min annealing temperature and 1 min at 72°C. A final extension step was performed for 7 min at 72°C. For all PCR-based procedures, amplification was done with a DNA thermal cycler 9600 (Applied Biosystems, Foster City, CA) in a final volume of 25 μL. A sample (3 μL) of each PCR product was electrophoresed in 1% (wt/vol) agarose (Gibco BRL) gel slabs (14 cm by 10 cm by 6 mm) in TAE buffer at 70 V. The gels were soaked in ethidium bromide solution (0.5 μg/mL) and the DNA was visualized in a transilluminator under UV light. The DNA molecular size marker used was a 1000-bp DNA ladder (Fermentas).

Saps gene library - The PCR products of family I genes (*C. albicans SAP1* to *SAP3* and *C. tropicalis SAPT4*) were purified using the PCR Purification Kit (Qiagen Inc, Valencia, CA), cloned into pCR®2.1-TOPO using the TOPO TA Cloning Kit and transformed into electrocompetent *Escherichia coli* Top-10 cells (Invitrogen). Recombinant plasmids were purified from overnight cultures by the alkaline lysis procedure (Sambrook & Rusell 2001) and digested with *Eco*RI restriction enzyme to release the inserts and generate *Eco*RI patterns. One clone per restriction fragment length polymorphism pattern was sequenced using M13 primers. The DNA sequences were determined using the Thermo Sequenase Cy5.5 Dye Terminator Cycle Sequencing Kit (Amersham Pharmacia Biotech, Piscataway, NJ) and an ABI PRISM 310 Genetic Analyzer (Perkin-Elmer Applied Biosystem, Inc, Boston, MA).

Multiple sequence alignment, similitude/identity percentages and phylogenetic analyses - The homologues of aspartyl proteases identified *in silico* and those of the previously characterized aspartyl proteases were aligned using ClustalX (Thomson et al. 1997) and MUSCLE in SeaView 2.4 program (Galtier et al. 1996) with default alignment parameter adjustments. A similitude and identity matrix were computed with the MatGAT4.50.2 software (Campanella et al. 2003). Phylogenetic analyses were performed with the MEGA4 program (Tamura et al. 2007) using neighbour-joining computed with the Poisson correction. To corroborate support for the branches on trees, bootstrap analysis (1,000 replicates) was performed.

Orthologous, superfamily, family and subfamily definitions - Traditionally, orthologues are defined as genes whose independent evolution reflects a speciation event (Fitch 1970) and their organisms of origin share a common ancestor. In this work, orthologues are defined as those genes contained in clusters of orthologous groups (COGs). Each COG consists of individual orthologous genes from two or more phylogenetic lineages or species; any two proteins from different lineages that belong to the same COG are orthologues (Tatusov et al. 2001). In addition, similitude/identity percentage, arrangement of motifs, synteny and phylogenetic grouping criteria were used in orthologue identification.

Protein superfamilies and families are defined as groups of related proteins that exhibit less than 50% and greater than 50% similarity, respectively (Dayhoff 1979). Subfamilies were defined as groups of proteins with at least 90% similarity and were often equivalent to COGs. However, in this work, other available criteria, such as functional similarity (substrate specificity, enzyme subcellular location, expression pattern etc.) and synteny were considered for subfamily detection. Genes lacking high similitude, common synteny, similar putative transcription factor binding sites, expression profile or physiological similitude were defined as paralogues.

RESULTS

The Sap protein motifs are summarized in Fig. 1. A total of 11 groups, each with different motif arrangements, were observed, but all analyzed proteins exhibited an amino terminal signal peptide, four cysteine residues and two aspartyl active site residues, the defining characteristics of Saps. PROSITE and Softberry software were unable to detect all aspartyl domains in the proteins, but local or manual alignments could detect them. In addition, some groups had serine, threonine or glutamine-rich regions. *C. albicans* Sap4-6 proteins have one or two putative RGD amino acid motifs. No internal sequence repeats were detected in any Sap with the TRUST program.

Construction of a phylogenetic tree of the Sap superfamily produced 12 families of genes from six *Candida* species: Sap1-3 and Sapt4, Sap4-6, Sapp1-3, Sapt1 and Sap8, Sapt2, Sapgu and Saplu, Sapt3, Sapp (2 families), Sap7, Sap10 and Sap9 (Fig. 2). *C. albicans* and *C. dubliniensis* had orthologous genes with high similarity (> 90%) organised in six families and clear synteny between *SAP1-4-SAP7-10* (Fig. 3), but *SAP5-6* were exclusive to *C. albicans.* A total of 11 *SAPT* genes, organised in seven families, were detected in *C. tropicalis*, including *SAPT4*, a possible orthologue of *SAP1* due to their 71.8% similitude, absence of RGD motif and partial synteny (Fig. 1, 3). A pair of *C. tropicalis* genes (CTRG02710/CTRG02711) organised in a cluster within the Sap10 family. In *C. parapsilosis*, 14 *SAPP* genes were recognised, including the previously described *SAPP1-SAPP3* genes. These *SAPP* genes organised into six families with no clear synteny with other spe-

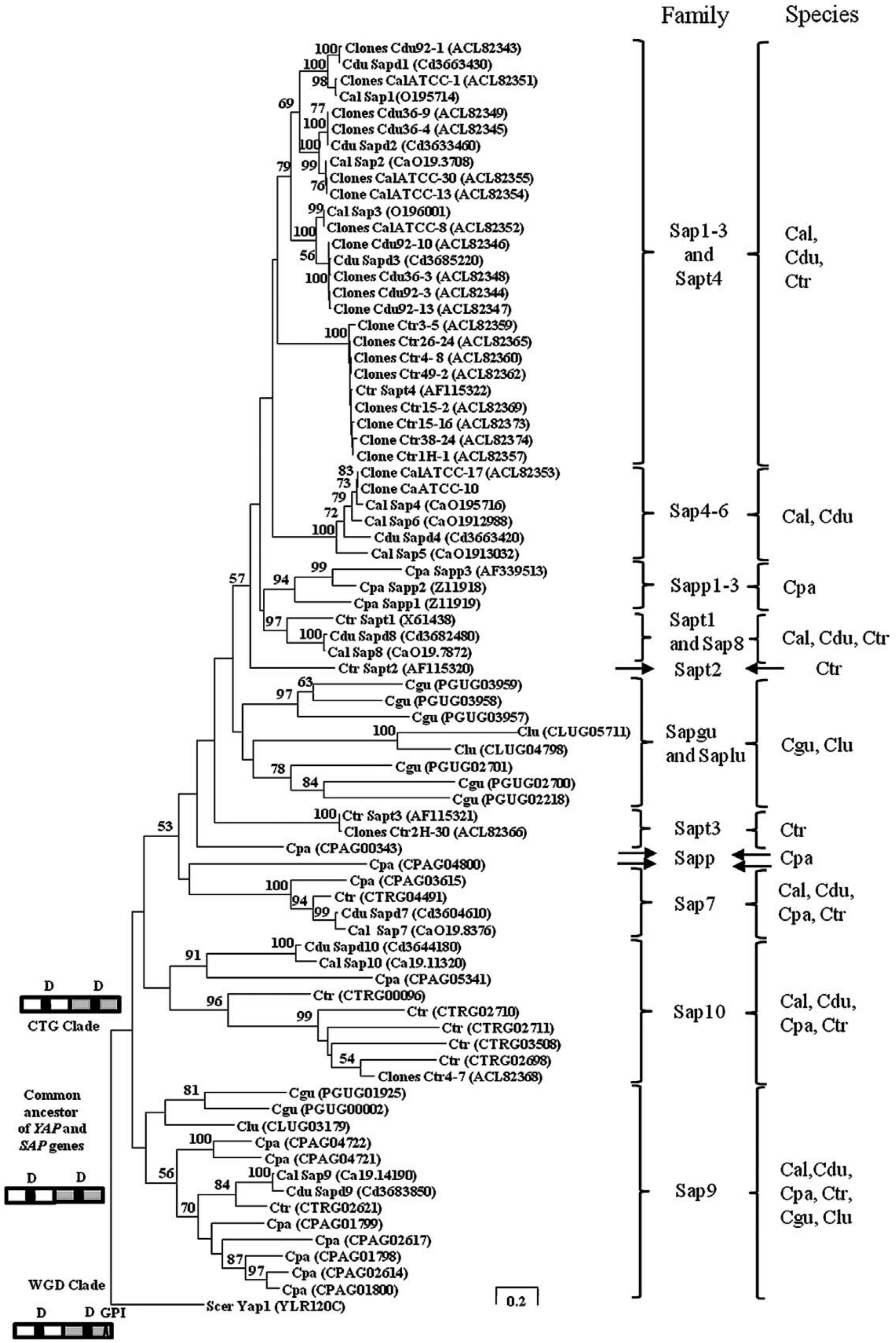

Fig. 2: neighbor-joining phylogenetic tree of secreted aspartyl proteinases (Saps) superfamily of opportunistic pathogenic *Candida* species. CTG clade, yeast that encodes CTG as serine instead of leucine. WGD clade, yeast that have undergone a genome duplication. Bootstrap values > 50% are on branches. Curly brackets and arrows indicate the Sap protein families defined by phylogenetic tree, similitude percentage (> 50%), synteny and motifs array. An example of each clone sequenced with different aminoacids sequence is showing. D: aspartic motif; GPI: glycosyl phosphatidyl binding site.

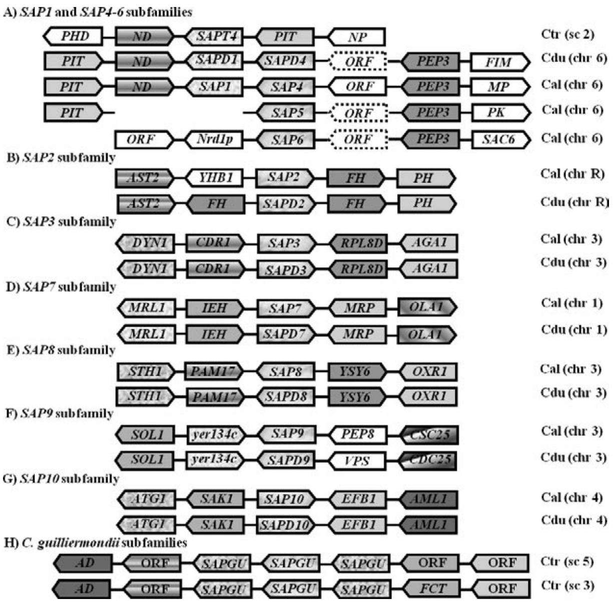
A) SAP1 and SAP4-6 subfamilies
PHD ND SAPT4 PIT NP Ctr (sc 2)
PIT ND SAPD1 SAPD4 ORF PEP3 FIM Cdu (chr 6)
PIT ND SAP1 SAP4 ORF PEP3 MP Cal (chr 6)
PIT SAP5 ORF PEP3 PK Cal (chr 6)
ORF Nrd1p SAP6 ORF PEP3 SAC6 Cal (chr 6)
B) SAP2 subfamily
AST2 YHB1 SAP2 FH PH Cal (chr R)
AST2 FH SAPD2 FH PH Cdu (chr R)
C) SAP3 subfamily
DYN1 CDR1 SAP3 RPL8D AGA1 Cal (chr 3)
DYN1 CDR1 SAPD3 RPL8D AGA1 Cdu (chr 3)
D) SAP7 subfamily
MRL1 IEH SAP7 MRP OLA1 Cal (chr 1)
MRL1 IEH SAPD7 MRP OLA1 Cdu (chr 1)
E) SAP8 subfamily
STH1 PAM17 SAP8 YSY6 OXR1 Cal (chr 3)
STH1 PAM17 SAPD8 YSY6 OXR1 Cdu (chr 3)
F) SAP9 subfamily
SOL1 yer134c SAP9 PEP8 CSC25 Cal (chr 3)
SOL1 yer134c SAPD9 VPS CDC25 Cdu (chr 3)
G) SAP10 subfamily
ATG1 SAK1 SAP10 EFB1 AML1 Cal (chr 4)
ATG1 SAK1 SAPD10 EFB1 AML1 Cdu (chr 4)
H) C. guilliermondii subfamilies
AD ORF SAPGU SAPGU SAPGU ORF ORF Ctr (sc 5)
AD ORF SAPGU SAPGU SAPGU FCT ORF Ctr (sc 3)

Fig. 3: synteny of *SAP* genes of *Candida albicans* (Cal), *Candida dubliniensis* (Cdu) and *Candida guilliermondii* (Cgu) genomes. A: *SAP1* and *SAP4-6* subfamily; B: *SAP2* subfamily; C: *SAP3* subfamily; D: *SAP7* subfamily; E: *SAP8* subfamily; F: *SAP9* subfamily; G: *SAP10* subfamily; H: *Candida guilliermondii* subfamilies. The genes in the rectangles encode the following proteins: *AD*: adenosine deaminase; *AGA1*: anchorage subunit; *AML1*: methyltransferase; *AST2*: [H+]ATPase of plasma membrane; *ATG1* and *SAK1*: serine/threonine kinases; *CDR1*: multidrug transporter of ATP-binding cassette (ABC); *CSC25*: guanyl-nucleotide exchange factor; *DYN1*: dynein heavy chain; *EFB1*: traslation elongation factor EF-1 beta; *FCT*: family of choline transport protein; *FH*: flavohemoprotein; *FIM*: fimbrin; *IEH*: protein exhibits interaction with ERG11 and Hsp82p; *MP*: monocarboxylate permeases; *MRL1*: membrane protein; *MRP*: mitochondrial ribosomal protein; *ND*: NADH dehydrogenase; *NP*: potential nucleoporin; *Nrd1p*: transcription is induced upon induction of hyphal growth; *OLA1*: P-loop ATPase; ORF: unknown predicted open reading frame; *OXR1*: oxidation resistance protein; *PAM17*: presequence translocase-associated motor subunit; *PEP3*: vacuolar membrane protein; *PEP8*: protein similar to *Saccharomyces cerevisiae* Pep8p; *PH*: phenol hydroxylase; *PHD*: potential PHD finger protein; *PIT*: phosphatidylinositol transfer protein; *PK:* protein kinase; *SAc6*: transcription is regulated by Nrg1p and Mig1p; *SOL1*: cell cycle regulator; *STH1*: ATPase component of the RSC chromatin remodeling complex; *RPL8B*: ribosomal protein; *yer134c*: uncharacterized protein yer134c homologue; *YHB1*: nitric oxide dioxygenase; *YSY6*: similar to the mammalian RAMP4 protein. Chr: chromosome; sc: supercontig.

cies genes and two pairs (CPAG01798/CPAG01799 and CPAG04721/CPAG04722) organised in clusters inside the Sap9 family. *C. guilliermondii* had eight *SAPGU* genes in the Saplu and Sapgu and Sap9 families, five of which (PGUG03957/PGUG03958/PGUG03959 and PGUG02700/PGUG02701) clustered into the Sapgu and Saplu family. *C. lusitaniae* had three *SAPLU* genes organised in two families: Saplu and Sapgu and Sap9. No *SAP* genes were detected in *C. glabrata*.

No new members of *SAP* families were detected in clinical strains of *Candida* spp. Representatives of each sequence from all sequenced *C. albicans*, *C. dubliniensis* and *C. tropicalis* clones were included in the phylogenetic tree and grouped with previous recognised subfamilies; no PCR products were obtained with the other species (data not shown). The rest of the clones were identical and are indicated with an equal symbol: CalATCC-1 (ACL82351) = CalATCC-3 = CalATCC-12 = Cal12-6 = Cal12-20 = Cal12-24; Cdu92-11 (ACL82343); Cdu36-9 (ACL82349); Cdu92-7 (ACL82350) = Cdu92-8 = Cdu92-17; Cdu36-4 (ACL82345) = Cdu36-7 = Cdu36-8; CalATCC-30 (ACL82355); Cal12-2 (ACL82356) = Cal12-8 = Cal12-24; CalATCC-13 (ACL82354); CalATCC-8 (ACL82352) = 22; Cdu92-10 (ACL82346); Cdu36-3 (ACL82348) = Cdu36-5 = Cdu36-6 = 13; Cdu92-3 (ACL82344) = Cdu92-9; Cdu92-13 (ACL82347); Ctr3-5 (ACL82359); Ctr26-24 (ACL82365) = Ctr26-25 = Ctr26-33 = Ctr385H-7; Ctr4-8 (ACL82360) = Ctr4-9 = Ctr49-2 = Ctr49-3 = Ctr49-5; Ctr56-5 (ACL82363); Ctr62-2 (ACL82364) = Ctr62-3; Ctr49-2 (ACL82362) = Ctr49-4; Ctr44-6 (ACL82361); Ctr56-2 (ACL82371); Ctr62-1 (ACL82372); Ctr15-2 (ACL82369); Ctr6H-24 (ACL82367); Ctr2H-9 (ACL82358); Ctr15-16 (ACL82373); Ctr38-24 (ACL82374); Ctr1H-1 (ACL82357) = 7; CalATCC-17 (ACL82353) = CalATCC-10; Ctr2H-30 (ACL82366) = Ctr25-35; Ctr4-7 (ACL82368) = Ctr44-7 (ACL82370) (Fig. 2).

All the Saps are predicted to have an intracellular localization, although their sizes (from 382-722 aa), pI values (from 3.0-8.3) and molecular masses (from 33.22-74.7 kDa) were widely variable. Putative transcription factor binding sites were also analyzed. An NIT2 site, commonly associated with genes induced by nitrogen, was detected in most *SAP* genes, but at different redundancy and relative positions. Other detected sites were GCR1, MIG1, complex HAP2/3/5, associated with activation and repression of glycolytic genes, and MATALPHA2 and MATA1, associated with yeast mating factors. Basically, all regulatory regions upstream of *SAP* genes were qualitatively and quantitatively different, except for those of *C. albicans*, *SAP1-4* and *C. dubliniensis SAP1-4 SAP7-10* orthologueous geness.

DISCUSSION

Phylogenetic studies of *Candida* spp., based on several concatenated alignments and complete genomes, reveal the polyphyletic character of the genus and define two clades in this genus of opportunistic pathogens. The CTG clade includes yeast that translate CTG as serine instead of leucine (*C. albicans*, *C. dubliniensis*, *C. tropicalis*, *C. parapsilosis*, and *C. lusitaniae*); the WGD clade contains yeast that have undergone a genome duplication (*S. cerevisiae* and *C. glabrata*) (Fitzpatrick et al. 2006). Thus, *C. dubliniensis* and *C. albicans* maintain a close phylogenetic relationship, suggesting that their speciation is a relatively recent event. The *C. tropicalis* and *C. parapsilosis* pair maintains a moderate relationship with *C. albicans*/*C. dubliniensis* pair, whereas the *C. guilliermondii* and *C. lusitaniae* pair is more remotely associated with *C. albicans*. Finally, *C. glabrata,* a pathogenic species, clusters among non-pathogenic yeast.

The Yap and Sap superfamilies of aspartyl proteases possess characteristic motifs defined in PROSITE and Softberry databases. In the present study, an expanded definition of aspartyl protease motifs based on new genomic data is proposed: [LIVMFGACTPSYF]-[LIVMTADNQSFH]-[LIVFSAE]-D-[STP]-[GS]-[STAV]-[STAPDENQY]-X-{LIVMFSTNCGQ}-[LIVMFGTAW].

Although we did not detect internal duplications in Sap proteins, it is likely that the Yap and Sap superfamilies are derived from an internal duplication event in an ancestor sequence. Because the similarity among Saps is relatively high, we proposed that Saps must be organised as a superfamily of paralogues with at least 12 paralogous families and nine orthologous subfamilies.

A total of 10 different Sap paralogues have been recognised in the *C. albicans* genome (Miyasaki et al. 1994, Monod et al. 1994, 1998, Felk et al. 2000) and they can be organised in six families: Sap1-3, Sap4-6, Sap7, Sap8, Sap9 and Sap10. In the *C. dubliniensis* genome, only one member of the Sap4-6 family is found (*SAPD4*) and the other *Candida* species have no Sap4-6 orthologues. Thus, *SAP5* and *SAP6* belong to a recently amplified gene family, exclusive to *C. albicans*, and may derive from a complete gene duplication event in an ancestral *SAP4* gene, *C. albicans* Sap4-6 proteins contain a functional RGD motif, a crucial tripeptide for cell surface receptor interaction. The presence of this motif might be explained by the fact that *SAP4-6* genes are expressed in the hyphae, which are more adherent than the yeast form (Hube et al. 1994, White & Agabian 1995).

C. albicans SAP1-3 and *SAP4-6* gene families are not pseudogenes that have subsisted to purifying selection and concerted evolution, but apparently have acquired new and specific functions in nitrogen metabolism, switching, and cell adherence (Cassone et al. 2002, Felk et al. 2002, Ripeau et al. 2002, Schofield et al. 2003). *C. tropicalis* has only one gene (*SAPT4*) related to the *C. albicans SAP1-3* and *SAP4-6* families. Based on similitude percentage, synteny and phylogenetic analyses, *SAPT4* is orthologous to *C. albicans SAP1* and *C. dubliniensis SAPD1.*

Why have *C. albicans/C. dubliniensis* undergone several events of *SAP* genes duplication? This question remains unanswered, but the diminished virulence of *C. albicans SAP*-null mutants suggests that a variety of *SAP* genes is necessary to host adaptation, surviving and virulence (Schaller et al. 2005). Similar gene duplication events in other Sap families can be observed in *C. tropicalis*, *C. lusitaniae, C. guilliermondii* and *C. parapsilosis* genomes.

SAP gene distribution among *C. dubliniensis* (Gilfillan et al. 1998), *C. tropicalis* (this work) and *C. albicans* populations seems to be universal (Bautista et al. 2003); however, a differential distribution of *SAP1-3* and *SAP4-6* gene families in *C. albicans* strains isolated from vaginal and systemic hematogenous infections has been observed (Kalkanci et al. 2005). The clonal structure of reproduction of *C. albicans* populations is typical of an asexual fungus and limits their capacity to generate diversity (Lott & Effat 2001, Tavanti et al. 2004); thus, gene duplication and rapid differentiation must be mechanisms that promote diversity. Nevertheless, it is very important to confirm or reject the universality of the *SAP* gene distributions in *C. albicans* populations. Each species has independently developed two types of gene duplication events, generating clusters of genes arranged in tandem and orphans located in remote genomic locations. Analyses of the distribution of *SAP* genes in the *Candida* species analyzed in this work confirmed the clustering of genes into previous recognised families, with exception of clones Ctr4-7 (ACL82368) and Ctr44-7 (ACL82370) that cluster with the Sap10 family.

Each member of *C. albicans SAP* gene family harbours regulatory regions with particular transcription factor binding site arrangements (this work) and transcriptional expression profiles (Monod et al. 1998, Naglik et al. 2003, 2004). Thus, each *C. albicans SAP* gene can be considered not only a duplicate functional gene, but also a different orthologue. To our knowledge, no *SAP* gene expression studies of *C. tropicalis, C. parapsilosis*, *C. guilliermondii* and *C. lusitaniae* have been performed to support this proposal.

It has been recently noted that *S. cerevisiae* is not a proper model for comparative studies of *C. albicans* (Odds 2008). Single gene inversions, chromosomal translocations, and long distance gene transpositions have generated two very distant species that diverged 700 million years ago (Seoighe et al. 2000). Thus, *C. albicans* should be used as a model for other *Candida* species. The exception is *C. glabrata,* which, despite the differences that have accumulated between it and *S. cerevisiae*, including kinetochore components of centromeres, expansion of sugar-utilization genes, loss of sex, loss of all active type I retroposons and genome reduction events (Stoyan & Carbon 2004, Dujon 2006), does not have a SAP superfamily. *C. glabrata* maintains at least eleven *YAP* genes, some of which are expressed in macrophage tissues (Kaur et al. 2007), whereas non-pathogenic *S. cerevisiae* possess only five members of the *YAP* superfamily (Krysan et al. 2005, Kaur et al. 2007). Because *C. glabrata* belongs to a non-pathogenic yeast clade, its virulence attributes may have evolved independently of those in the CTG clade.

The transition from a commensal life style to opportunistic state seems to correlate aspartyl protease gene expansion. The niche of *Candida* species, frequently found as normal biota in humans, has selected against pathogenic strains (Calderone 2002). However, at least in *C. glabrata*, *C. tropicalis*, *C. parapsilosis* and *C. lusitaniae,* some virulence tasks are evolving (Odds 2008). We hypothesize that non-pathogenic yeast species contain fewer *SAP* and *YAP* genes than their close, opportunistic pathogen relatives, which will only be revealed after their genomes are sequenced and released.

ACKNOWLEDGMENTS

To Bernard Dujon, for donating strains.

REFERENCES

Bautista-Muñoz C, Boldo XM, Villa-Tanaca L, Hernández-Rodríguez C 2003. Identification of *Candida* spp. by randomly amplified polymorphic DNA and differentiation between

C. albicans and *C. dubliniensis* by direct PCR methods. *J Clin Microbiol 41*: 414-420.

Calderone RA 2002. *Candida* and candidiasis, ASM Press, Washington DC, p. 15.

Campanella JJ, Bitincka L, Smalley J 2003. MatGAT: an application that generates similarity/identity matrices using protein or DNA sequences. *BMC Bioinformatics 4*: 29.

Cassone A, Tacconelli E, De Bernardis F, Tumbarello M, Torosantucci A, Chiani P, Cauda R 2002. Antiretroviral therapy with protease inhibitors has an early, immune reconstitution-independent beneficial effect on *Candida* virulence and oral candidiasis in human immunodeficiency virus-infected subjects. *J Infect Dis 185*: 188-195.

Dayhoff MO 1979. *Atlas of protein sequence and structure*, vol. 5, Suppl. 3, National Biomedical Research Foundation, Silver Springs, Maryland, p. 353-358.

Dujon B 2006. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet 22*: 375-387.

Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A 2002. The PROSITE database. *Nucleic Acids Res 30*: 235-238.

Felk A, Kretschmar M, Albrecht A, Schaller M, Beinhauer S, Nichterlein T, D Sanglard D, Korting HC, Schäfer W, Hube B 2002. *Candida albicans* hyphal formation and the expression of the Efg1-regulated proteinases Sap4 to Sap6 are required for the invasion of parenchymal organs. *Infect Immun 70*: 3689-3700.

Felk A, Schafer W, Hube B 2000. *Candida albicans* secretory aspartic proteinase (*SAP10*) Gene Accession number AF146440. Available from: http://www.ncbi.nlm.nih.gov/protein/28863901.

Fitch WM 1970. Distinguishing homologous from analogous proteins. *Syst Zool* 19: 99-113.

Fitzpatrick D, Logue ME, Stajich JE, Butler G 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol Biol 6*: 99.

Galtier N, Gouy M, Gautier C 1996. SeaView and Phylo_win, two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci 12*: 543-548.

Gilfillan GD, Sullivan DJ, Haynes K, Parkinson T, Coleman DC, Gow NAR 1998. *Candida dubliniensis*: phylogeny and putative virulence factors. *Microbiology 144*: 829-838.

Hoffman CS, Winston F 1987. A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of *Escherichia coli*. *Gene 57*: 267-272.

Hoyer LL, Fundyga R, Hecht JE, Kapteyn JC, Klis FM, Arnold J 2001. Characterization of agglutinin-like sequence genes from non-*albicans Candida* and phylogenetic analysis of the ALS family. *Genetics 157*: 1555-1567.

Hube B, Monod M, Schofield DA, Brown AJP, Gow NAR 1994. Expression of seven members of the gene family encoding secretory aspartyl proteinases in *Candida albicans*. *Mol Microbiol 14*: 87-99.

Hube B, Naglik JR 2002. Extracellular hydrolases. In RA Calderone, *Candida and candidiasis*, ASM Press, Washington DC, p. 107-122.

Kalkanci A, Bozdayi G, Biri A, Kustimur S 2005. Distribution of secreted aspartyl proteinase using a polymerase chain reaction assay with *SAP* specific primers in *Candida albicans* isolates. *Folia Microbiol 50*: 409-413.

Kaur R, Ma B, Cormack BP 2007. A family of glycosylphosphatidylinositol-linked aspartyl proteases is required for virulence of *Candida glabrata*. *Proc Natl Acad Sci USA 104*: 7628-7633.

Krysan DJ, Ting EL, Abeijon C, Kroos L, Fuller RS 2005. Yapsins are a family of aspartyl protease required for cell wall integrity in *Saccharomyces cerevisiae*. *Eukaryot Cell 4*: 1364-1374.

Lott TJ, Effat MM 2001. Evidence for a more recently evolved clade within a *Candida albicans* North American population. *Microbiology 147*: 1687-1692.

Miyasaki SH, White TC, Agabian N 1994. A fourth secreted aspartyl proteinase gene (*SAP4*) and a *CARE2* repetitive element are located upstream of the *SAP1* gene in *Candida albicans*. *J Bacteriol 176*: 1702-1710.

Monod M, Hube B, Hess D, Sanglard D 1998. Differential regulation of *SAP8* and *SAP9*, which encode two new members of the secreted aspartic proteinase family in *Candida albicans*. *Microbiology 144*: 2731-2737.

Monod M, Togni G, Hube B, Sanglard D 1994. Multiplicity of genes encoding secreted aspartic proteinases in *Candida* species. *Mol Microbiol 13*: 357-368.

Moran GP, Derek JS, Coleman DC 2002. Emergence of Non-*Candida albicans Candida* species as pathogen. In RA Calderone, *Candida and candidiasis*, ASM Press, Washington DC, p. 37.

Naglik J, Albrecht A, Bader O, Hube B 2004. *Candida albicans* proteinases and host/pathogen Interactions. *Cell Microbiol 6*: 915-926.

Naglik JR, Challacombe SJ, Hube B 2003. *Candida albicans* secreted aspartyl proteinases in virulence and pathogenesis. *Microbiol Mol Biol Rev 67*: 400-428.

Odds FC 2008. Evolution of pathogenic *Candida* species. In F Baquero, C Nombela, GH Cassell, JA Gutiérrez-Fuentes, *Evolutionary biology of bacterial and fungal pathogens,* ASM Press, Washington DC, p. 567-580.

Pfaller MA, Diekema DJ 2007. Epidemiology of invasive candidiasis: a persistent public health problem. *Clin Microbiol Rev 20*: 133-163.

Pichova I, Pavlickova L, Dostal J, Dolejsi E, Hruskova-Heidingsfeldova O, Weber J, Ruml T, Soucek M 2001. Secreted aspartic proteases of *Candida albicans*, *Candida tropicalis*, *Candida parapsilosis* and *Candida lusitaniae*. Inhibition with peptidomimetic inhibitors. *Eur J Biochem 268*: 2669-2677.

Quandt K, French K, Kara H, Wingender E, Werner T 1995. MatInd and MatInspector -new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res 23*: 4878-4884.

Ripeau JS, Fiorillo M, Aumont F, Belhumeur P, de Repentigny L 2002. Evidence for differential expression of *Candida albicans* virulence genes during oral infection in intact and human immunodeficiency virus type 1-transgenic mice. *J Infect Dis 185*: 1094-1102.

Sambrook J, Rusell DW 2001. *Molecular cloning: a laboratory manual*, 3rd ed., Cold Spring Harbor Laboratory Press, New York, section 1.31-1.32.

Schaller M, Korting HC, Borelli C, Hamm G, Hube B 2005. *Candida albicans*-secreted aspartic proteinases modify the epithelial cytokine response in an *in vitro* model of vaginal candidiasis. *Infect Immun 73*: 2758-2765.

Schofield DA, Westwater C, Warner T, Nicholas PJ, Paulling EE, Balish E 2003. Hydrolytic gene expression during oroesophageal and gastric candidiasis in immunocompetent and immunodeficient gnotobiotic mice. *J Infect Dis 188*: 591-599.

Seoighe C, Federspiel N, Jones T, Hansen N, Bivolarovic V, Surzycki R, Tamse R, Komp C, Huizar L, Davis RW, Scherer S, Tait E, Shaw DJ, Harris D, Murphy L, Oliver K, Taylor K, Rajandream MA, Barrell BG, Wolfe KH 2000. Prevalence of small inversions in yeast gene order evolution. *Proc Natl Acad Sci USA 97*: 14433-14437.

Stehr F, Felk A, Kretschmar M, Schaller M, Schafer W, Hube B 2000. Extracellular hydrolytic enzymes and their relevance during *Candida albicans* infections. *Mycoses 43*: 17-21.

Stoyan T, Carbon J 2004. Inner kinetochore of the pathogenic yeast *Candida glabrata*. *Eukaryot Cell 3*: 1154-1163.

Szklarczyk R, Heringa J 2004. Tracking repeats using significance and transitivity. *Bioinformatics 20*: 311-317.

Tamura K, Dudley J, Nei M, Kumar S 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol 24*: 1596-1599.

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res 29*: 22-28.

Tavanti A, Gow NA, Maiden MC, Odds FC, Shaw DJ 2004. Genetic evidence for recombination in *Candida albicans* based on haplotype analysis. *Fungal Genet Biol 41*: 553-562.

Thomson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG 1997. The ClustalX Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res 24*: 4876-4882.

White TC, Agabian N 1995. *Candida albicans* secreted aspartyl proteinases: isoenzyme pattern is determined by cell type and levels are determined by environmental actors. *J Bacteriol 177*: 5215-5221.

Zaugg C, Borg-Von Zepelin M, Reichard U, Sanglard D, Monod M 2001. Secreted aspartic proteinase family of *Candida tropicalis*. *Infect Immun 69*: 405-412.