Nucleic Acids Research, 2005, Vol. 33, Database issue ***D353–D357***
doi:10.1093/nar/gki124

CandidaDB: a genome database for *Candida albicans* pathogenomics

C. d'Enfert*, S. Goyard, S. Rodriguez-Arnaveilhe, L. Frangeul[1], L. Jones[2], F. Tekaia[3], O. Bader[4], Antje Albrecht[4], L. Castillo[5], A. Dominguez[6], J. F. Ernst[7], C. Fradin[4], C. Gaillardin[8], S. Garcia-Sanchez, P. de Groot[9], B. Hube[4], F. M. Klis[9], S. Krishnamurthy[7], D. Kunze[4], M.-C. Lopez[6], A. Mavor[10], N. Martin[6], I. Moszer[1], D. Onésime[8], J. Perez Martin[11], R. Sentandreu[5], E. Valentin[5] and A. J. P. Brown[10]

Unité Postulante Biologie et Pathogénicité Fongiques, INRA USC 2019, [1]Génopole Plate-forme Intégration et Analyse Génomiques, [2]Groupe Logiciels et Banques de Données and [3]Unité de Génétique Moléculaire des Levures, CNRS URA 2171, Département Structure et Dynamique des Génomes, Institut Pasteur, Paris, France, [4]Robert Koch Institute, NG4, Berlin, Germany, [5]University of Valencia, Burjassot, Spain, [6]Universidad de Salamanca, Salamanca, Spain, [7]Heinrich-Heine-Universität, Düsseldorf, Germany, [8]Laboratoire de Génétique Moléculaire et Cellulaire, INA-PG-INRA-CNRS, Thiverval-Grignon, France, [9]Universiteit van Amsterdam, Swammerdam Institute for Life Sciences, Amsterdam, The Netherlands, [10]Aberdeen University, Aberdeen, UK and [11]Centro Nacional de Biotecnologia-CSIC, Madrid, Spain

Received July 30, 2004; Revised October 4, 2004; Accepted October 21, 2004

ABSTRACT

CandidaDB is a database dedicated to the genome of the most prevalent systemic fungal pathogen of humans, *Candida albicans*. CandidaDB is based on an annotation of the Stanford Genome Technology Center *C.albicans* genome sequence data by the European Galar Fungail Consortium. CandidaDB Release 2.0 (June 2004) contains information pertaining to Assembly 19 of the genome of *C.albicans* strain SC5314. The current release contains 6244 annotated entries corresponding to 130 tRNA genes and 5917 protein-coding genes. For these, it provides tentative functional assignments along with numerous pre-run analyses that can assist the researcher in the evaluation of gene function for the purpose of specific or large-scale analysis. CandidaDB is based on GenoList, a generic relational data schema and a World Wide Web interface that has been adapted to the handling of eukaryotic genomes. The interface allows users to browse easily through genome data and retrieve information. CandidaDB also provides more elaborate tools, such as pattern searching, that are tightly connected to the overall browsing system. As the *C.albicans* genome is diploid and still incompletely assembled, CandidaDB provides tools to browse the genome by individual supercontigs and to examine information about allelic sequences obtained from complementary contigs. CandidaDB is accessible at http://genolist.pasteur.fr/CandidaDB.

INTRODUCTION

Candida sp. are ubiquitous yeasts commonly isolated from the environment. Among the 200 species described, a few are commensals of humans and of several animal species (1). *Candida* sp. are also opportunistic pathogens in humans, being responsible for superficial as well as life-threatening systemic infections, mainly in hospitalized individuals (2). Among *Candida* sp., *Candida albicans* is responsible for the majority of all forms of candidiasis (3). Consequently, in recent years *C.albicans* has been the focus of a broad range of studies aimed at understanding its pathogenesis and population dynamics, identifying targets for the development of novel antifungals and eventually restricting the incidence of *Candida* infections in hospital settings (4).

*To whom correspondence should be addressed at Unité Postulante Biologie et Pathogénicité Fongiques, INRA USC 2019, Institut Pasteur, 25 rue du Docteur Roux, 75015 Paris, France. Tel: +33 1 40 61 32 57; Fax: + 33 1 45 68 89 38; Email: denfert@pasteur.fr
Present addresses:
S. Rodriguez-Arnaveilhe, Aventis Pharma, LGI-Bioinformatics, Vitry s/Seine, France
S. Garcia-Sanchez, Department of Biotechnology, NEIKER, Vitoria-Gazteiz, Spain

As *C.albicans* is an obligate diploid, forward genetics is tedious in this species and post-genomic approaches are especially important for the exploration of the molecular mechanisms that underlie *C.albicans* pathogenesis (4). In this context, whole-genome shotgun sequencing of *C.albicans* has been undertaken, resulting in the release of successive assemblies of the *C.albicans* diploid genome sequence and associated sets of open reading frames (ORFs) of more than 100 amino acids (5). The latest assembly, Assembly 19, is distributed over 412 supercontigs, of which 266 constitute a reference haploid genome of 14 855 kb and 146 constitute allelic counterparts of supercontigs included in the reference haploid genome (5). The reference haploid genome contains 7677 ORFs of 100 codons or longer, and a reduced set of 6419 ORFs has been derived by eliminating the smaller of a pair of ORFs that overlap by more than 50% (5). However, a detailed annotation of these ORF sets has not been provided, nor a convenient interface that would allow researchers to query the *C.albicans* genome sequence, the gene set or the protein set in multiple ways.

The principal aim of CandidaDB is to provide a complete annotated genomic sequence of *C.albicans* SC5314. CandidaDB is based on GenoList, a generic relational data schema and a user-friendly World Wide Web interface allowing rapid searching and visualization of genomic features (6). The current release of CandidaDB, launched in June 2004, provides tentative functional assignments for 130 tRNA genes and 5918 protein-coding genes identified in Assembly 19. It also provides data for the rapid evaluation of *C.albicans* protein function, intracellular location, topology and protein family membership.

SOURCE DATA AND METHODS

Source data and identification of annotation-relevant ORFs

Nucleotide sequence data for Assemblies 5, 6 and 19 of the *C.albicans* strain SC5314 genome sequence were retrieved from the Stanford Genome Technology Center (SGTC) website (http://www-sequence.stanford.edu/group/candida/). The current release of CandidaDB is based on Assembly 19, composed of a haploid supercontig set (contigs 19-831–19-10 262), here referred to as the haploid set, and an allelic supercontig set (contigs 19-20001–19-20 161), here referred to as the allelic set (5). The CAAT-box software package (7) was used to identify annotation-relevant ORFs in Assemblies 5, 6 and 19. Assembly-specific GeneMark matrices (8) were built from a set including ORFs longer than 300 codons and a set with all intergenic regions obtained after subtraction of ORFs larger than 80 codons. ORFs longer than 150 codons were systematically retained for further annotation and assigned a reference number of the format IPF*n.i* (where IPF stands for Individual Protein File, *n* is an integer rank specific to the IPF and *i* corresponds to the number of times the IPF has been modified between Assemblies 5, 6 and 19 of the *C.albicans* genome sequence). ORFs with a length between 40 and 150 codons were also selected and assigned an IPF number provided that they have a GeneMark coding function of more than 0.5 over their whole length, do not overlap with a larger IPF on a different reading frame and show a significant match in the database of non-redundant proteins available from the NCBI (BLASTP *E*-value < $1e^{-3}$) (9,10).

Annotation of Assembly 6 of the *C.albicans* genome

A total of 8890 IPFs were identified in Assembly 6. These IPFs were subjected to manual annotation using the annotation interface of the CAAT-box software package (7). Functions were assigned on the basis of published data when available or similarity to proteins of known function, the latter being explicitly indicated in the function field. The standard convention for naming *C.albicans* genes was used (http://hypha.stanford.edu/Nomenclature.shtml). Only genes that have already been characterized or can be postulated to encode a functional homologue of the most closely related *Saccharomyces cerevisiae* gene were named according to this convention. Genes that did not meet these criteria were assigned a formal gene name of the format IPF*n*.

Several tags have been added to gene names to take into account the occurrence of frame-shifts and contig breaks still present in the assemblies of the *C.albicans* genome sequence (Supplementary Table 1S). Assembly 6 shows sequence redundancy because of the diploid nature of *C.albicans* (5). Therefore, in order to identify duplicated ORFs, each IPF was checked against a database including all IPFs, using BLASTP (10). Artefactual duplications were confirmed by comparing 5′- and 3′-non-coding regions using BLASTN (10) and one of the duplicated IPFs was assigned as FALSORF, leading to a non-redundant gene set being included in CandidaDB. This analysis also resulted in the identification of protein families encoded by the *C.albicans* genome. Families not identified previously in *C.albicans* and likely to have emerged through species-specific amplification were designated with a gene name of the format IFX*n* (where IF stands for IPF Family, X is a letter specific to the gene family and *n* is an integer; Supplementary Table 2S). This process resulted in a non-redundant set of 6165 *C.albicans* full-length or partial proteins. This set was used to build the first release of CandidaDB (CandidaDB.v1, January 2002) in which all proteins were assigned an entry number of the type CA*nnnn*.

Annotation of the haploid set of Assembly 19 of the *C.albicans* genome

Annotation data for Assembly 6 were used to re-annotate a group of 11 616 ORFs identified from the Assembly 19 haploid set using both the strategy outlined above and data available from the SGTC (5). Re-annotation was performed using the annotation tool Artemis (11). Chromosome assignments were obtained from the Biotechnology Research Institute—National Research Council Canada website (http://candida.bri.nrc.ca/candida/contigs/index.html). The tRNAs were predicted using tRNAScan-SE [(12); http://www.genetics.wustl.edu/eddy/tRNAscan-SE/]. Intron–exon structures were predicted on the basis of similarity to other known proteins or the lack of a start codon within the identified ORF. This procedure resulted in 6244 annotated non-redundant features corresponding to 5918 protein-coding genes and 130 tRNA genes.

Altogether 9552 ORFs were identified in the Assembly 19 allelic set using the procedures described above. The haploid and allelic sets of ORFs were compared using reciprocal BLASTP (10) in order to correlate the 6114 *C.albicans* protein features to their allelic counterparts. A reciprocal comparison to the *S.cerevisiae* proteome was performed using data available at the *Saccharomyces* Genome Database (13) to

identify potential direct orthologues in the two organisms. The protein features were checked against common protein motifs and families using the Pfam database (14) and were analyzed for the occurrence of signal sequences and membrane-spanning domains using SignalP (15) and TMHMM 2.0, respectively (16).

Finally, using BLASTP (10), each of the 6114 *C.albicans* protein sequences was checked against a database including all of these sequences. Sequences showing a BLASTP *E*-value of $<10^{-10}$ (10,17) were considered significantly similar and were, on the basis of this criterion, considered non-unique in the set of the surveyed *C.albicans* sequences. This set of non-unique proteins was partitioned into disjoint subsets of related proteins denoted P*n.m* (where *n* is the number of members in the partition and *m* an arbitrary partition order). Partitions were subsequently analyzed using the MAST system (18) to define amino acid motifs shared among the different members of each partition. Independently, non-unique proteins were clustered using the Markov Clustering (MCL) algorithm (19). The MCL clustering was performed using −log(BLASTP *E*-values) and an inflation index $I = 3.0$. Clusters were denoted C*p.q*, where *p* is the number of members in the cluster and *q* is an arbitrary cluster order.

Implementation of CandidaDB

CandidaDB is based on a general data frame, called GenoList (6). GenoList databases are based on a relational data schema and a World Wide Web interface originally developed for the handling of bacterial genomes and adapted to the handling of eukaryotic genomes. The database structure and data are run on a mainframe computer using the Sybase® SQL DBMS. The Sybase implementation is used to generate dynamic HTML pages through a CGI script written in the C/C++ and Perl programming languages.

THE CandidaDB WORLD WIDE WEB INTERFACE

The World Wide Web interface of CandidaDB uses a presentation common to all GenoList databases that facilitates an exploration and in-depth analysis of genome information through typical questions asked by the biologist (6). The front page is divided into three HTML frames. The left-hand frame makes the most common queries accessible including searches by gene names and synonyms, contig location and keywords. More sophisticated queries are also accessible including classical sequence analysis tools [BLAST (10) and FASTA (20)] and a pattern-searching tool that allows the user to search for degenerate patterns in nucleic acid and protein sequences. The upper-right frame shows either a list generated from simple or complex queries or a graphical and dynamic representation of a contig region. Nucleic acid and protein sequences which correspond to the region or those genes that are included in it can be exported. Finally, the lower-right frame presents detailed information about the gene of interest selected from the upper frame. In the case of CandidaDB, this includes, when available, contig and chromosome assignment, gene name and function, synonyms (ORF 19 and ORF 6 numbers defined at the SGTC, IPF number), information about the closest *S.cerevisiae* homologue and an HTML link to the relevant web page at SGD (13). Further inclusions are HTML links to relevant web pages at the SGTC (5), MIPS (21) or Genbank and to allelic sequences in Assembly 19 when available, summarized results of SignalP analysis, a summary of the membrane-spanning domain prediction with a link to relevant data, Pfam domains and links to relevant web pages at the EBI, contribution to a protein partition and link to MAST analysis, and regularly updated Smith and Waterman scanning reports of a range of protein databases [nrprot (9), *S.cerevisiae* (13), *Schizosaccharomyces pombe* (22) and *C.albicans*]. From this frame, it is also possible to access the neighboring genome region and to retrieve nucleotide and amino acid sequences together with adjacent sequences.

CONTENT OF THE CURRENT RELEASE

CandidaDB Release 2.0 (June 2004) contains information pertaining to Assembly 19 of the genome of *C.albicans* strain SC5314 (5). The current release contains 6244 annotated entries corresponding to 130 tRNA genes and 5918 protein-coding genes. In contrast to the previously reported set of *C.albicans* proteins (5), the data available in CandidaDB include proteins shorter than 100 amino acids that (i) were identified through the Genmark coding function of the ORFs or through sequence similarity with known proteins and (ii) correspond to genuine small proteins or partial ORFs resulting from natural or artefactual frame-shifts in the haploid set of the current assembly of the *C.albicans* genome. Using this information, it has been possible to identify *C.albicans* proteins by peptide mass fingerprinting with an 89% success rate. This rate is equivalent to that obtained in equivalent *S.cerevisiae* proteomics experiments (23). Hence, CandidaDB has become an invaluable tool for *C.albicans* proteomics research, as well as for transcript profiling studies (24–27).

Table 1 provides a summary of the characteristics of the protein-coding features defined on the basis of various

Table 1. Summary of the protein features in CandidaDB

Characteristics of proteins	Number of entries (%)
Total	6114 (100)
With an allelic counterpart	5315 (86.9)
With PFAM match(es)	3712 (60.7)
Within a partition[a]	2103 (34.4)
With a tentative signal peptide[b]	694 (11.4)
With at least one membrane-spanning domain[c]	1238 (20.3)
With a homologue[d]	5611 (91.8)
With a homologue in all three phylogenetic domains	1781 (29.1)
With a homologue only in the eukaryotic domain	2606 (42.6)
With a homologue only in Ascomycetes	1712 (28.0)
Unique to *Candida albicans*	608 (9.9)
With a direct orthologue[a] in *Saccharomyces cerevisiae*	3809 (62.3)

[a]See Source Data and Methods for definition.
[b]As defined using SignalP (16).
[c]As defined using TMHMM 2.0 (15).
[d]*C.albicans* proteins were compared with 540 687 proteins identified within the proteomes of eukaryotic ($n = 27$), bacterial ($n = 53$) and archeal ($n = 19$) fully sequenced species and to UniProt (31) filtered from any of these previous 540 687 sequences. Homology was deemed significant when the BLASTP *E*-value was $<10^{-10}$ for eukaryotic proteins and $<10^{-4}$ for most bacterial or archeal proteins.

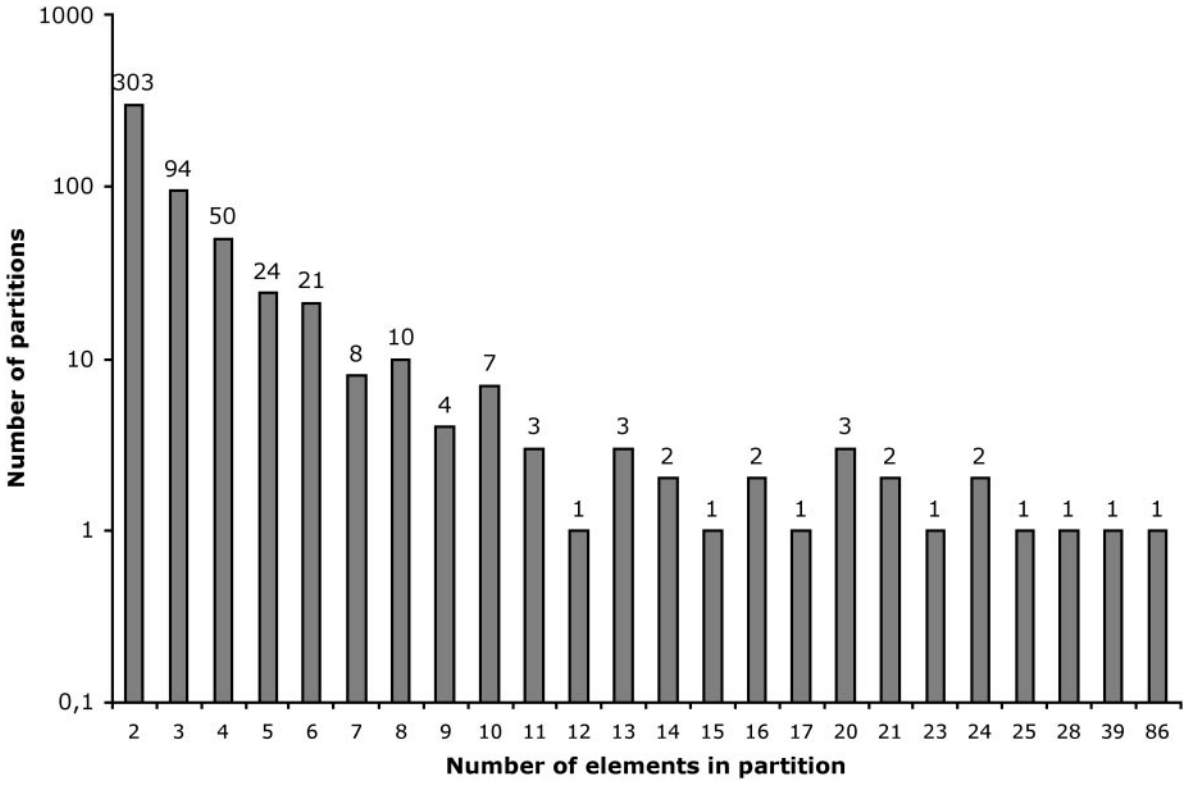

Figure 1. Overall genome redundancy as deduced from partition analysis. The total number of partitions according to their size is shown. Partitions were established as described in the Source Data and Methods section. Proteins within a partition have at least one reciprocal BLASTP link with another protein in the partition with an *E*-value $<e^{-10}$. The resulting overall genome redundancy defined as the ratio (%) of the number of proteins belonging to partitions over the total number of proteins is 34.4%.

analyses: protein localization and topology, identification of common protein motifs (see also Supplementary Table 3S) and assignment to protein partitions identified within the *C.albicans* proteome (Figure 1 and see also Supplementary Tables 4S and 5S). Partitioning constitutes a powerful subdivision of related proteins but might include diverse proteins that are related solely through a single domain. Hence, Markov Clustering (MCL) was used as a parallel and complementary method since it favors closely related proteins. Supplementary Table 6S shows that partitions and MCL clusters were concordant for most families with sometimes artificially distinct clusters of homogeneously related proteins. However, partitions including numerous members were subdivided into homogeneous clusters. Some partitions appear to correspond to gene families that may have arisen through species-specific expansion (Supplementary Table 2S). The existence of these gene families has been predicted on the basis of the sequence similarity of their members, as was the case originally for the SAP and LIP gene families (28,29), for example, and it will be interesting to investigate the roles of these families in the pathobiology of this fungus.

FUTURE DIRECTIONS AND CONCLUSION

The CandidaDB genome database for *C.albicans* pathogenomics was launched in January 2002 through the concerted efforts of nine European groups constituting the Galar Fungail network (http://www.pasteur.fr/Galar_Fungail/) that annotated Assembly 6 of the *C.albicans* genome released earlier by the Stanford Genome Technology Center. Other *C.albicans* genomics databases do exist (http://web.ahc.umn.edu/biodata/candida/; http://agabian.ucsf.edu/). Nevertheless, since CandidaDB was first released, it has attracted strong interest worldwide, averaging 30 000 hits per month. It is now recognized as a reference database for *C.albicans* studies (4,30). It has served to build and is connected to the *C.albicans* PEDANT database available from MIPS (21). The latest release of CandidaDB launched in June 2004 provides an up-to-date annotation of Assembly 19 of the *C.albicans* genome sequence with numerous pre-run analyses that can assist the researcher in the evaluation of gene function for the purpose of specific or large-scale analysis. Following the release of Assembly 19 by the SGTC, a community effort was launched in order to reach a homogeneous annotation (Braun *et al*., manuscript in preparation; http://candida.bri.nrc.ca/candida/index.cfm) that should soon become available through the Candida Genome Database (http://www.candidagenome.org/). This annotation will be integrated within CandidaDB in the near future with the aim of providing complementary information to CGD. Also, the structure of CandidaDB, which can handle individual contigs, is perfectly suited to the integration of future assemblies of the *C.albicans* genome. Furthermore, in the future, we aim to upgrade the database model of CandidaDB with a view to integrating the numerous genomes that are becoming available for yeasts and filamentous ascomycetes.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

Sequence data from *C.albicans* were obtained from the Stanford Genome Technology Center (http://www.sequence.stanford.edu/group/candida). Sequencing of *C.albicans* was accomplished with the support of the NIDR and the Burroughs Wellcome Fund. This work was supported by grants from the European Commission (QLK2-2000-00795; MCRTN-CT-2003-504148; 'Galar Fungail Consortium') to A.J.P.B., C.E., A.D., J.E., C.G., B.H., F.M.K., J.P.M. and R.S. and the Ministère de la Recherche et de la Technologie (PRFMMIP 'Réseau Infections Fongiques') to C.E. and C.G. F.T. was supported by the Institut Pasteur Strategic Horizontal Program on *Anopheles gambiae*. N.M. was supported by a fellowship of the Junta de Castilla y Leon and by grants DGCYT (PM-98-0317 and BIO 2002-02124) to A.D. R.S. was supported in part by grants from the Spanish Ministerio de Ciencia y Tecnología (BMC2003-01023) and Agència Valenciana de Ciència i Tecnologia de la Generalitat Valenciana (Grupos 03/187).

REFERENCES

1. Calderone,R. (2002) Taxonomy and biology of *Candida*. In Calderone,R. (ed.), *Candida and Candidiasis*. ASM Press, Washington DC, pp. 307–325.
2. Wenzel,R.P. (1995) Nosocomial candidemia: risk factors and attributable mortality. *Clin. Infect. Dis.*, **20**, 1531–1534.
3. Pfaller,M.A., Jones,R.N., Doern,G.V., Sader,H.S., Messer,S.A., Houston,A., Coffman,S. and Hollis,R.J. (2000) Bloodstream infections due to *Candida* species: SENTRY antimicrobial surveillance program in North America and Latin America, 1997–1998. *Antimicrob. Agents Chemother.*, **44**, 747–751.
4. Berman,J. and Sudbery,P.E. (2002) *Candida albicans*: a molecular revolution built on lessons from budding yeast. *Nature Rev. Genet.*, **3**, 918–930.
5. Jones,T., Federspiel,N.A., Chibana,H., Dungan,J., Kalman,S., Magee,B.B., Newport,G., Thorstenson,Y.R., Agabian,N., Magee,P.T. *et al.* (2004) The diploid genome sequence of *Candida albicans*. *Proc. Natl Acad. Sci. USA*, **101**, 7329–7334.
6. Moszer,I., Jones,L.M., Moreira,S., Fabry,C. and Danchin,A. (2002) SubtiList: the reference database for the *Bacillus subtilis* genome. *Nucleic Acids Res.*, **30**, 62–65.
7. Frangeul,L., Glaser,P., Rusniok,C., Buchrieser,C., Duchaud,E., Dehoux,P. and Kunst,F. (2004) CAAT-Box, contigs-assembly and annotation tool-box for genome sequencing projects. *Bioinformatics*, **20**, 790–797.
8. Lukashin,A.V. and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, **26**, 1107–1115.
9. Wheeler,D.L., Church,D.M., Edgar,R., Federhen,S., Helmberg,W., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E. *et al.* (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.*, **32**, D35–D40.
10. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
11. Rutherford,K., Parkhill,J., Crook,J., Horsnell,T., Rice,P., Rajandream,M.A. and Barrell,B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.
12. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
13. Christie,K.R., Weng,S., Balakrishnan,R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman,J.E. *et al.* (2004) *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res.*, **32**, D311–D314.
14. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
15. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
16. Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.
17. Tekaia,F., Blandin,G., Malpertuy,A., Llorente,B., Durrens,P., Toffano-Nioche,C., Ozier-Kalogeropoulos,O., Bon,E., Gaillardin,C., Aigle,M. *et al.* (2000) Genomic exploration of the hemiascomycetous yeasts: 3. Methods and strategies used for sequence analysis and annotation. *FEBS Lett.*, **487**, 17–30.
18. Bailey,T.L. and Gribskov,M. (1998) Combining evidence using *p*-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.
19. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
20. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
21. Frishman,D., Mokrejs,M., Kosykh,D., Kastenmuller,G., Kolesov,G., Zubrzycki,I., Gruber,C., Geier,B., Kaps,A., Albermann,K. *et al.* (2003) The PEDANT genome database. *Nucleic Acids Res.*, **31**, 207–211.
22. Hertz-Fowler,C., Peacock,C.S., Wood,V., Aslett,M., Kerhornou,A., Mooney,P., Tivey,A., Berriman,M., Hall,N., Rutherford,K. *et al.* (2004) GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res.*, **32** (Database issue), D339–D343.
23. Yin,Z., Stead,D., Selway,L., Walker,J., Riba-Garcia,I., Mclnerney,T., Gaskell,S., Oliver,S.G., Cash,P. and Brown,A.J.P. (2004) Proteomic response to amino acid starvation in *Candida albicans* and *Saccharomyces cerevisiae*. *Proteomics*, **4**, 2425–2436.
24. Garcia-Sanchez,S., Aubert,S., Iraqui,I., Janbon,G., Ghigo,J.M. and d'Enfert,C. (2004) *Candida albicans* biofilms: a developmental state associated with specific and stable gene expression patterns. *Eukaryot. Cell*, **3**, 536–545.
25. Fradin,C., Kretschmar,M., Nichterlein,T., Gaillardin,C., d'Enfert,C. and Hube,B. (2003) Stage-specific gene expression of *Candida albicans* in human blood. *Mol. Microbiol.*, **47**, 1523–1543.
26. Rogers,P.D. and Barker,K.S. (2003) Genome-wide expression profile analysis reveals coordinately regulated genes associated with stepwise acquisition of azole resistance in *Candida albicans* clinical isolates. *Antimicrob. Agents Chemother.*, **47**, 1220–1227.
27. Karababa,M., Coste,A.T., Rognon,B., Bille,J. and Sanglard,D. (2004) Comparison of gene expression profiles of *Candida albicans* azole-resistant clinical isolates and laboratory strains exposed to drugs inducing multidrug transporters. *Antimicrob. Agents Chemother.*, **48**, 3064–3079.
28. Hube,B., Stehr,F., Bossenz,M., Mazur,A., Kretschmar,M. and Schafer,W. (2000) Secreted lipases of *Candida albicans*: cloning, characterisation and expression analysis of a new gene family with at least ten members. *Arch. Microbiol.*, **174**, 362–374.
29. Monod,M., Togni,G., Hube,B. and Sanglard,D. (1994) Multiplicity of genes encoding secreted aspartic proteinases in *Candida* species. *Mol. Microbiol.*, **13**, 357–368.
30. Pitarch,A., Sanchez,M., Nombela,C. and Gil,C. (2003) Analysis of the *Candida albicans proteome*. II. Protein information technology on the Net (update 2002). *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, **787**, 129–148.
31. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*, **32**, D115–D119.