Comparative genomics using *Candida albicans* DNA microarrays reveals absence and divergence of virulence-associated genes in *Candida dubliniensis*

Gary Moran,¹ Cheryl Stokes,¹ Sascha Thewes,² Bernhard Hube,² David C. Coleman¹ and Derek Sullivan¹

Correspondence

Derek Sullivan

derek.sullivan@dental.tcd.ie

¹Microbiology Research Unit, Department of Oral Surgery, Oral Medicine and Oral Pathology, School of Dental Science, University of Dublin, Trinity College, Dublin 2, Republic of Ireland

²Robert Koch Institut, Nordufer 20, D-13353 Berlin, Germany

Candida dubliniensis is a pathogenic yeast species closely related to *Candida albicans*. However, it is less frequently associated with human disease and displays reduced virulence in animal models of infection. Here comparative genomic hybridization was used in order to assess why *C. dubliniensis* is apparently less virulent than *C. albicans*. In these experiments the genomes of the two species were compared by co-hybridizing *C. albicans* microarrays with fluorescently labelled *C. albicans* and *C. dubliniensis* genomic DNA. *C. dubliniensis* genomic DNA was found to hybridize reproducibly to 95.6% of *C. albicans* gene-specific sequences, indicating a significant degree of nucleotide sequence homology (>60%) in these sequences. The remaining 4.4% of sequences (representing 247 genes) gave *C. albicans*/*C. dubliniensis* normalized fluorescent signal ratios that indicated significant sequence divergence (<60% homology) or absence in *C. dubliniensis*. Sequence divergence was identified in several genes (confirmed by Southern blot analysis and sequence analysis of PCR products) with putative virulence functions, including the gene encoding the hypha-specific human transglutaminase substrate Hwp1p. Poor hybridization of *C. dubliniensis* genomic DNA to the array sequences for the secreted aspartyl proteinase-encoding gene *SAP5* also led to the finding that *SAP5* was absent in *C. dubliniensis* and that this species possesses only one gene homologous to *SAP4* and *SAP6* of *C. albicans*. In addition, divergence and absence of sequences in several gene families was identified, including a family of *HYR1*-like GPI-anchored proteins, a family of genes homologous to a putative transcriptional activator (*CTA2*) and several *ALS* genes. This study has confirmed the close relatedness of *C. albicans* and *C. dubliniensis* and has identified a subset of unique *C. albicans* genes that may contribute to the increased prevalence and virulence of this species.

Received 7 April 2004

Revised 22 July 2004

Accepted 27 July 2004

INTRODUCTION

Candida dubliniensis is associated with oral candidosis in the HIV-infected population in which this species was first identified in 1995 (Jabra-Rizk *et al.*, 2001; Sullivan *et al.*, 1995, 2004). Recently, several studies have also identified *C. dubliniensis* as a cause of oral disease in diabetic and cancer patients (Sebti *et al.*, 2001; Willis *et al.*, 2000). However, the closely related species *Candida albicans*

appears to be more successful than *C. dubliniensis* as a commensal of the human oral cavity in healthy individuals, as determined by standard oral swab sampling methods (Coleman *et al.*, 1997). In addition, the incidence of *C. dubliniensis* isolation from blood cultures is extremely low compared to *C. albicans* (Kibbler *et al.*, 2003; Pfaller & Dieker, 2004). In a recent study of *Candida* species recovered from blood cultures in six sentinel hospitals in England and Wales between 1997 and 1999, *C. dubliniensis* was isolated from only 2% of samples compared to 65% for *C. albicans*. Two studies have also demonstrated that *C. dubliniensis* is less virulent than *C. albicans* in a murine model of systemic candidosis (Gilfillan *et al.*, 1998; Vilela *et al.*, 2002). The reason for the apparent difference in virulence between the two species is unknown as they are

Abbreviations: CGH, comparative genomic hybridization; GPI, glycosyl-phosphatidylinositol.

The GenBank/EMBL/DDBJ accession numbers for the sequences reported in this paper are AJ634664–AJ634672, AJ634675–AJ634676 and AJ634382.

phenotypically very similar and seem to share many of the traits traditionally associated with virulence in *C. albicans*. In particular both species have the ability to form true hyphae, to adhere to human epithelium and to produce secreted aspartyl proteinases (Gilfillan *et al.*, 1998; Hannula *et al.*, 2000; Vilela *et al.*, 2002). However, *C. dubliniensis* does not form hyphae as rapidly as *C. albicans* in response to shifts in pH/temperature or when incubated in serum (Gilfillan *et al.*, 1998). In contrast, when cultured on Staib agar or Pal's agar *C. dubliniensis* forms abundant hyphae, pseudohyphae and chlamydospores, whereas *C. albicans* remains in the yeast phase (Al Mosaïd *et al.*, 2001, 2003). *C. dubliniensis* also seems to be more sensitive to environmental stress, such as elevated temperature and NaCl concentration (Alves *et al.*, 2002; Pinjon *et al.*, 1998).

Comparative genomic hybridization (CGH) studies with DNA microarrays provide a rapid and cost-effective method for obtaining informative data about uncharacterized genomes and have been used extensively to compare gene content in prokaryotic and eukaryotic micro-organisms (Daran-Lapujade *et al.*, 2003; Dong *et al.*, 2001; Murray *et al.*, 2001). The completion of the *C. albicans* genome project and the availability of *C. albicans* DNA microarrays now enables genomes of different strains of *C. albicans* and closely related species, such as *C. dubliniensis*, to be compared. In the present study, CGH was performed between *C. albicans* and *C. dubliniensis* using *C. albicans* DNA microarrays in order to identify genomic differences that might account for the difference in virulence between *C. albicans* and *C. dubliniensis*. This approach was deemed feasible as all *C. dubliniensis* genes analysed to date share more than 90% identity at the nucleotide sequence level with the orthologous *C. albicans* genes. Total genomic DNA from *C. albicans* and *C. dubliniensis* was co-hybridized to *C. albicans* DNA microarrays and the relative hybridization efficiency of *C. dubliniensis* and *C. albicans* DNA to each gene-specific spot was compared. This approach allowed us to identify the presence of thousands of *C. albicans*-homologous genes in *C. dubliniensis* without the need for sequence analysis and has guided us towards genes which are highly divergent or even absent from *C. dubliniensis*. We anticipate that this collection of *C. albicans*-specific sequences may contain genes that contribute to the observed differences in virulence and epidemiology between these two organisms.

METHODS

Candida strains and culture conditions. *C. albicans* strain SC5314 was used as a control in all CGH experiments using Eurogentec *C. albicans* DNA microarrays. *C. dubliniensis* strains used in this study included the *C. dubliniensis* type strain CD36 (American Type Culture Collection reference MYA-178, British National Collection of Pathogenic Fungi reference NCPF3949), which is a representative of *C. dubliniensis* Cd25 fingerprint group I (genotype 1), and *C. dubliniensis* strain CD514, a strain representative of Cd25 fingerprint group II (genotype 3) (Gee *et al.*, 2002). Strains were routinely grown on potato dextrose agar (PDA; Oxoid)

medium, pH 5.6, at 37°C. For liquid culture, cells were grown in yeast extract-peptone-dextrose (YEPD) broth, also at 37°C (Gallagher *et al.*, 1992).

Chemicals, enzymes and radioisotopes. All chemicals used were of molecular biology grade and were purchased from Sigma-Aldrich. Molecular biology enzymes and kits were purchased from Promega or New England Biolabs unless otherwise indicated. Cy5 and Cy3 dUTP were purchased from Amersham Biosciences Europe. Supplies of [α -³²P]dATP (6000 Ci mmol⁻¹, 220 TBq mmol⁻¹) were purchased from NEN Life Sciences.

DNA microarrays. *C. albicans* DNA microarrays used in this study were constructed by Eurogentec based on the Galar Fungail consortium's annotation of the *C. albicans* SC5314 genome sequence in the CandidaDB (http://www.pasteur.fr/Galar_Fungail/CandidaDB/). This annotation was produced based on the genome sequence released by the Stanford Genome Technology Center. Each glass slide microarray contained sequences corresponding to 6039 ORFs (98% of annotated genes) that were approximately 300 bp in length and spotted in duplicate.

Genomic DNA preparation. High-molecular-mass total genomic DNA was recovered from *Candida* strains by organic extraction following digestion of the cell wall with Zymolyase 20T (Seikagaku) and proteinase K treatment (Roche Diagnostics) as described by Gallagher *et al.* (1992).

Genomic DNA labelling and microarray hybridization. For DNA labelling experiments with Cy5 dUTP and Cy3 dUTP, total genomic DNA (2 µg) was first fragmented by either restriction endonuclease digestion or sonication. For restriction endonuclease digests, two 1 µg aliquots of DNA were separately digested with *Tru*II or *Rsa*I (Fermentas). These digests were then heat inactivated, extracted once with a mixture of phenol/chloroform/isoamyl alcohol (25:24:1, by vol.) and ethanol precipitated. The two separate aliquots of digested DNA were combined to give a mixture of *Tru*II and *Rsa*I fragments (50–4000 bp) for labelling. Alternatively, separate DNA samples were prepared for labelling by sonication using a Sonoplus HD70 sonicator (Bandelin Electronic) at 75% power for 30 cycles producing DNA fragments ranging from 500 to 5000 bp.

Each labelling reaction was carried out with 2 µg of either sheared or digested genomic DNA using the RadPrime random priming labelling system (Invitrogen) incorporating Cy5 dUTP or Cy3 dUTP into *C. albicans* SC5314 or *C. dubliniensis* DNA fragments. After labelling, reaction products were purified with Nucleospin PCR clean up columns (Macherey-Nagel) and concentrated to a final volume of <5 µl with a Microcon YM-30 column (Millipore). Cy5-labelled and Cy3-labelled reactions were mixed together in DIG EasyHyb buffer (Roche Diagnostics) to a final volume of 60 µl for hybridization. The mixture was denatured at 98°C for 5 min then chilled on ice. Microarray slides (Eurogentec) were placed in a hybridization chamber (Corning), covered with a glass LifterSlip (Erie Scientific Company) and the labelling reaction was carefully applied at the edges of the slide. The chamber was sealed and incubated in the dark at 42°C for 16–18 h. Slides were washed at high stringency at room temperature as follows: (i) 5 min in 1 × SSC, 0.03% (w/v) SDS, (ii) 5 min in 0.2 × SSC and (iii) 5 min in 0.05 × SSC. Following washing, slides were dried thoroughly by centrifugation at low speed for 5 min in a 50 ml disposable plastic tube (Greiner Bio-One) and scanned immediately.

Each of the hybridizations performed using digested DNA and sheared DNA was carried out on two separate occasions. A third sheared DNA hybridization was performed with the Cy3 and Cy5

dyes swapped. One additional hybridization was performed between sheared Cy5-labelled *C. albicans* DNA and sheared Cy3-labelled *C. dubliniensis* CD514 DNA.

Data analysis. DNA microarray slides were scanned with the GenePix 4000B scanner (Axon Instruments). Data were extracted from scanned images using the GenePix Pro 4.1.1.4 software package (Axon Instruments). Data normalization and subsequent analysis was carried out with the GeneSpring 6.1 software package (Silicon Genetics). Hybridization data from each DNA 'spot' on the slide were only included for analysis if the control (*C. albicans*) channel signal was above local background plus two standard deviations. Signal intensities in both channels were background corrected. Measurements were normalized across the whole chip by dividing each measurement by the median of all measurements taken for that chip. A normalized fluorescence ratio value was determined for each spot by dividing the *C. albicans* control channel normalized signal by the *C. dubliniensis* normalized signal values. The $\log_2$ value of each ratio was determined and the $\log_2$ ratios of duplicate spots were averaged. The significance of normalized ratios of <1 was determined in replicate experiments using Student's *t* test. The raw data from all of the experiments has been submitted in a MIAME-compliant format to the ArrayExpress database at the European Bioinformatics Institute (accession code E-MEXP-99).

The relationship between $\log_2$ ratio values and nucleotide sequence homology was determined by linear regression analysis using Prism 4.0 (GraphPad Software). For this analysis, nucleotide sequence homology between the array printed *C. albicans* sequences (~ 300 bp) and the corresponding region of available homologous *C. dubliniensis* sequences was determined using DNA Strider 3.1 software. Sequences

used in this analysis included the available *C. dubliniensis* gene sequences from GenBank and PCR-amplified sequences described here (Table 1). Sequences were included for analysis only when a minimum of 100 bp of uninterrupted sequence could be aligned. Gaps of over 50 bp in length were excluded from homology calculations. On the basis of this analysis, we chose genes with normalized ratios of <0.5 ($P<0.05$) in replicate dye-swap experiments with sheared genomic DNA for further study.

Larger sequence alignments (>500 bp) described in the results were carried out using the CLUSTAL_W software package (Higgins & Sharp, 1988).

Southern hybridization. Southern hybridization analysis was carried out as described previously (Moran *et al.*, 1998; Southern, 1975) using DNA sequences labelled with [α - 32 P]dATP by random primer labelling (Prime-a-Gene system; Promega) or using DIG-labelled probes incorporating DIG-11-dUTP (Roche Diagnostics) during PCR amplification as described by the manufacturers. In order to identify homologues of individual ORFs in *C. dubliniensis*, the entire *C. albicans* ORF was used in labelling reactions. In the case of gene families (e.g. *CTA2*, *IFA*, *IPF*), probes were designed by identifying the most conserved region of the ORFs in CLUSTAL_W alignments. These gene family probes represented $>35\%$ of the entire ORF and constituted a different region of the ORF to that present on the microarray. The post-hybridization washes were performed at low stringency (60 °C with $0.5 \times$ SSC, 0.1%, w/v, SDS) unless otherwise indicated.

PCR amplification of *C. dubliniensis* genome sequences. PCR amplification from the *C. dubliniensis* genomic DNA template

Table 1. Percentage nucleotide sequence homology of *C. dubliniensis* genes to corresponding *C. albicans* Eurogentec microarray sequences

Gene	Percentage homology	Normalized ratio*	GenBank accession no.	Reference
<i>ACT1</i>	98.8	>1.5	AJ236897	Donnelly <i>et al.</i> (1999)
<i>URA3</i>	93.0	>1.5	AJ302032	Staib <i>et al.</i> (2001)
<i>MDR1</i>	92.0	1.4	AJ227752	Moran <i>et al.</i> (1998)
<i>CDR1</i>	91.9	1.16	AJ439073	Moran <i>et al.</i> (2002)
<i>CDR2</i>	91.0	>1.5	AJ439075	Moran <i>et al.</i> (2002)
<i>ERG3</i>	90.1	1.17	AJ421248	Pinjon <i>et al.</i> (2003)
<i>ERG11</i>	90.4	>1.5	AY034876	Perea <i>et al.</i> (2002)
<i>PHR2</i>	90.5	1.40	AF184908	Kurzai <i>et al.</i> (1999)
<i>PHR1</i>	88.0	>1.5	AF184907	Kurzai <i>et al.</i> (1999)
<i>SAP4</i>	86.0	0.55	AJ634382	This study
<i>SAP2</i>	83.6	1.18	AJ634672	This study
<i>CZF1</i>	80.0	0.53	AJ634475	This study
<i>IPF8147</i>	79.6	0.29	AJ634476	This study
<i>IPF1873</i>	75.0	0.24	AJ634664	This study
<i>RVS161</i>	75.0	0.30	AJ634665	This study
<i>IPF16104</i>	74.0	0.25	AJ634666	This study
<i>IFA8</i>	70.4	0.36	AJ634667	This study
<i>IPF9787</i>	70.0	0.26	AJ634668	This study
<i>IPF3448</i>	69.6	0.42	AJ634669	This study
<i>SSN6</i>	68.0	0.30	AJ634670	This study
<i>IPF2057</i>	59.0	0.17	AJ634671	This study

*Mean ratio of normalized fluorescence values of *C. albicans* and *C. dubliniensis* hybridizing DNA at each gene-specific spot. Values above 1.5 are quoted as >1.5 for the purposes of regression analysis (Fig. 1).

Table 2. Sequences of oligonucleotide primers used in this study

Primer name*	Sequence (5'–3')	Nucleotide coordinates†
CRH12F	ACTGGAATGGGAACTGAAC	+1159 to +1178
CRH12R	CACAACACTACTGAAAGATG	+1476 to +1457
1760F	CAGAAATCTGGTATTGACAC	+712 to +731
1760R	TCTATATCCATATCGCATTC	+1121 to +1102
11560F	CAATCTAAACCTCCATCAGG	+451 to +470
11560R	AGGGGAATTACTAATGACTC	+821 to +812
8147F	CCCTACTACTTCATCATCAC	+405 to +424
8147R	AATTAATCCCTTCAGAATAC	+769 to +750
CHS5F	AGGAAACAGATATTGTTGAG	+1055 to +1074
CHS5R	TTCTGTAGATGTTGGCTCAG	+1476 to +1457
IPF3448F	GGAGTATTTGGAGACCCAAG	+492 to +511
1PF3448R	TGGCATTGTTCTTCACCAAC	+868 to +849
BIO3F	CAGGAACATGCTGGTATCTG	+791 to +809
BIO3R	ACCCAAACACCTAAATCGAC	+1187 to +1168
HIK1.3F	AAAAGTCTAACCCAATTGAC	+443 to +462
HIK1.3R	TTCATAATTGTAGTGATCG	+843 to +824
HIK1.5F	AAAACGTTAGCCGTCAAAGC	+1753 to +1772
HIK1.5R	TCTAATATTAGATGGCGACA	+2202 to +2183
HIT1F	TGTCCTAAATGTTCAATTGC	+49 to +58
HIT1R	ATTCATCAATTCAAGACATC	+444 to +425
HNM4F	TAGGTGGAGAACCAATTGTG	+887 to +906
HNM4R	AAAGCACACCCAAAGGACAG	+1350 to +1331
HNM4A	ATGTCTTATCAAATTACAGA	+1 to +20
HNM4B	ATCAATATTATGATCATAGC	+1497 to +1478
CZF1F	ATCTCAACCTTTGTATTCTG	+657 to +676
CZF1R	TTTCGTCATCCTTTTGATCC	+1087 to +1068
8627F	AATCAACAACGTGTTAATCG	+1249 to +1268
8627R	TATTGATAAATTACCATGAG	+1671 to +1652
15920F	CTACCACAATAAACAAACAG	+1433 to +1452
15920R	ACTTTGTTGTAATGATTGAG	+1875 to +1856
2057F	ACTACTGTTCTGCTGCTAC	+967 to +986
2057R	CCCTTGATATCAACAATGTC	+1385 to +1366
2971F	AATTGCTAAACAAGGAGATC	+927 to +956
2971R	TATCTTTTCCTTGTTCTAAC	+1372 to +1352
DAL52F	TGTTGTTGGATGTATTATCC	+1119 to +1138
DAL52R	CATCTTCACTATTACGTGCT	+1558 to +1539
FUR4F	AGGGTTCATTCTGCTAATTG	+1281 to +1300
FUR4R	GATACCATCATGCACTTCAT	+1662 to +1643
FUR4A	AAAAGTTGGTTGTCTCCATG	+26 to +45
FUR4B	TCTTCGAAGTACAGATTCC	+1702 to +1683
HNM3F	ATTTTGACTGGTATCGTTTG	+976 to +995
HNM3R	GATACATAATTCATGTTGGT	+1409 to +1390
HNM3A	ATGGATACCAAGTATGAAAAC	+1 to +21
HNM3B	ATTCAGTCTTATTGGCAATG	+1495 to +1476
BIO4F	TGGAAGCCCATTCAAACAGG	+91 to +110
BIO4R	CCACGGTTCCTCAAATGCTC	+467 to +448
OPT1F	GGTAAAGTTTCTTCAATGC	+1843 to +1862
OPT1R	AGTCGGTGTTTGTTAACTC	+2239 to +2220
OPTA	AAAATAAGGGCAGTAATTAG	+7 to +26
OPTB	TCAATATGTGTCTGATATTG	+2317 to +2298
IFF2F	AATGGTTCCTGGTAATGGATC	+3340 to +3359
IFF2R	CAAGAAAACAACAACCATTG	+3728 to +3747
4805F	AAAACCTTATGCTTTTGAG	+4069 to +4088
4805R	AATTCAATACCTAACATACC	+4418 to +4399

Table 2. cont.

Primer name*	Sequence (5'–3')	Nucleotide coordinates†
IPF257F	TACCAAACATCATGTTTCACG	+58 to +78
IPF257R	AATGACTGATTCATAGTA	+402 to +383
1873F	GAATTTGCTAAACGAATCGG	+433 to +452
1873R	AGATGACTGTTTAAATCGAG	+886 to +869
RSV161F	ACAATCGCTCTACTCGAATG	+246 to +265
RSV161R	CAAGTACTGTTGGATCTGTG	+690 to +671
RRN3F	CGCCGCATTTTCAGGCATTGC	+1248 to +1267
RRN3R	CTATATATCGTCTTCACTATC	+1671 to +1651
13135F	TAATTGTTTTGATAGCAATG	+218 to +238
13135R	GATATTGATGAATCATTAGC	+590 to +571
9787F	AAGAACAGTTGGATTCAGAG	+989 to +1008
9787R	AATTGATCATTTCTTGGACG	+1349 to +1330
SSN6F	ACCAGTTAACCACCTGTTG	+2817 to +2836
SSN6R	TCATCATAATTTTCATCTTC	+3233 to +3116
16104F	AACTCTAAACAATTGTTGAC	+1741 to +1760
16104R	TCATACAAGATTCATTTTGG	+2172 to +2153
IFA1F	GCGAATTCAGTATGGACTTTATGTATC	+1279 to +1297
IFA1R	GCGAATTCCTACTCCATTATCCCGATC	+1969 to +1951
HYR1F	GAATTCAGGTTTCCATGGTGAT	+95 to +117
HYR1R	GAATTCAGCAGTGAAGATGATTG	+1200 to +1183
IFF1F	GCGAATTCGCTTAATTCTGTCTTAGC	+1 to +20
IFF1R	GCGAATTCATAAACTAGGATTATAACC	+844 to +826
SAP4-6F	GCGAATTCATATCTTGAGTGTCTTGC	+14 to +32
SAP4-6R	GCGAATTCCTTGCCCTTGTCATACC	+763 to +745
InvSAPF	GCGAATTCACGCAACAGCAAGAACACTC	+40 to +21
InvSAPR	GCGAATTCCTACTGAGTCTATGTATGAC	+623 to +642
CdSAP4F	CGTCTAGAGGAGGAACCTTGACGATGT	–226 to –207
CdSAP4R	CGCTCGAGCTTTTCATTTCTAGGCATATG	+1463 to +1444
RBTF2	GCGAATTCGAAGAATTAAGTAACGATGGT	+694 to +714
RBTR2	GCTCTAGAAGAAGTGACTGAAGTAGAATC	+1410 to +1390
HWP1F	CGGAATTCGATGAGATTATCAACTGCT	–14 to +5
HWP1R	CGGAATTCGGAATTAGATCAAGAATGCAG	+1908 to +1889
HWP2R	GGAATTCCTAGGATTGTCACAAGG	+223 to +210
APL6F	CGGAATTCGGAATACAAAGATGTTTC	+1678 to +1693
CTA2F	GCGAATTCATGCCAGAAAACCTCCAAAC	+1 to +20
CTA26R	GCGAATTCCTTCGTTTACGTGGTTGGTG	+781 to +751

*Primer names refer to gene annotations in CandidaDB (<http://genolist.pasteur.fr/CandidaDB/>).

†Nucleotide coordinates are given for each gene, where +1 refers to the first base of the ATG start codon. All coordinates are for *C. albicans* genes with the exception of InvSAPF/R and CdSAPF/R, which refer to the coordinates of the *CdSAP4* gene, and HWP2R, which refer to the *CdHWP1* gene.

was carried out as described previously (Moran *et al.*, 1998, 2002). Oligonucleotide primers used in this study were synthesized by Sigma-Genosys (Table 2). PCR-amplified DNA fragments were sequenced where indicated using the dideoxy chain-termination method by Lark Technologies.

RESULTS

Comparative genomic microarray hybridization

To label *Candida* chromosomal DNA efficiently by random priming with Cy3 or Cy5 dUTP, it was first necessary to

fragment the chromosomal DNA. Two DNA fragmentation methods, sonication and restriction endonuclease digestion, were compared in order to determine if either labelling method introduced artifacts into the microarray results. In replicate experiments involving *C. albicans* SC5314 genomic DNA labelled with Cy3 dUTP following restriction endonuclease or sonication, over 98 % of spots gave fluorescent signals that were two standard deviations above background. Using the same criteria, *C. dubliniensis* CD36 genomic DNA prepared by either sonication or restriction endonuclease digestion hybridized to at least 95 % of gene-specific spots in replicate hybridizations. The dataset chosen

for analysis here was generated with sonicated genomic DNA from *C. albicans* and *C. dubliniensis* labelled with either Cy5 dUTP or Cy3 dUTP in dye-swapped replicate experiments. In these hybridizations, 5931 duplicate spots (98 %) exhibited signals in the *C. albicans* channel two standard deviations above background and were included for analysis.

Relative hybridization efficiency of co-hybridized *C. albicans* SC5314 and *C. dubliniensis* CD36 genomic DNA to the microarrays was assessed by determining a normalized ratio of *C. albicans* and *C. dubliniensis* signal intensities at each spot included for analysis (Cy5/Cy3 normalized ratios). In dye-swapped replicate experiments, 96 % of these normalized ratios differed by less than twofold.

In order to investigate whether a relationship existed between the strength of hybridization of *C. dubliniensis* DNA to the array and the degree of nucleotide homology between the corresponding *C. albicans* and *C. dubliniensis* sequences, we plotted $\log_2$ ratio values versus percentage nucleotide sequence homology. We determined the nucleotide sequence homology between the *C. albicans* probe sequences present on the array and the corresponding sequences of 11 *C. dubliniensis* genes sequences available in GenBank (Table 1). We also attempted to PCR amplify sequences from *C. dubliniensis* using *C. albicans*-specific oligonucleotide primers (Table 2) corresponding to 35 genes which hybridized poorly with *C. albicans* genomic DNA (normalized ratios ranging from 0.17 to 0.53) on the microarray. Under low-stringency conditions, 10 of these PCR primer sets yielded PCR amplification products with sequences homologous to the corresponding *C. albicans* gene (nucleotide sequence homology range 59–80 %, Table 1).

Fig. 1. Standard curve used to determine the relationship between percentage nucleotide sequence homology of *C. albicans* SC5314 and *C. dubliniensis* CD36 sequences and normalized fluorescence ratio. GenBank sequences for 11 *C. dubliniensis* genes of known homology and 10 novel PCR-amplified sequences were included. The mean of the $\log_2$ ratio values (Table 1) for each gene was plotted against percentage nucleotide sequence homology. Linear regression analysis was used to predict the best-fitting line. The r^2 value was calculated using Prism 4.0.

Table 1). We plotted the percentage nucleotide sequence homology between the 11 GenBank and 10 PCR-amplified sequences and their *C. albicans* homologues versus the $\log_2$ ratio values from the 21 DNA spots representing these genes on the array (Fig. 1). Linear regression analysis was used to generate a best-fit-line from the dataset, which demonstrated a relationship between nucleotide sequence homology and $\log_2$ ratio ($r^2=0.80$, $P<0.0001$). The 12 most homologous nucleotide sequences (80–98.8 % identity), including housekeeping genes such as *ACT1*, *URA3* and *ERG11*, all had normalized ratio values of >0.5 . The remaining nine genes formed a second group with intermediate sequence homologies of 59–79 %. These nine sequences possessed normalized ratios ranging from 0.17 to 0.46. Based on this analysis we categorized genes into three homology groups based on normalized ratio values: (I) a high-homology group (normalized ratio >0.5 ; >80 % nucleotide sequence homology), (II) a medium-homology group (normalized ratio 0.25–0.5; 60–80 % nucleotide sequence homology) and (III) sequences that possessed low homology or were possibly absent in *C. dubliniensis* (normalized ratio <0.25 ; <60 % nucleotide sequence homology).

In total, 677 sequence-specific spots exhibited normalized ratios below 0.5 ($P<0.05$) in replicate dye-swap array experiments with sheared genomic DNA. From this group, 418 sequences were classified as likely to possess intermediate nucleotide sequence homology (60–80 %). The remaining 259 sequences (representing 4.4 % of the spots analysed) gave normalized ratios of <0.25 and were predicted to possess low nucleotide sequence homology (<60 %) or were possibly absent in *C. dubliniensis*.

Categorization of divergent genes

We decided to examine the group of sequences predicted to possess low nucleotide sequence homology in more detail as these genes were most likely to be functionally different or possibly even absent in *C. dubliniensis*. This group of 259 sequences was found to correspond to 247 genes (Table 3) since 12 genes were found to be represented on the array by two different duplicate spots. However, 134 (54 %) of these were hypothetical genes with no homology to genes of known function. Within this group, 43 genes could be classified as conserved hypothetical genes due to significant homology to other hypothetical genes in GenBank or CandidaDB (Table 4). Twenty-one sequences (8.5 %) were found to have homology to genes encoding *C. albicans* retrotransposon elements, including transposases and reverse transcriptases. Two sequences identified, IPF17727 and HOK, are contained in the *C. albicans*-specific fingerprinting probes 27A and Ca3 (Soll, 2000). IPF17727 is a region of the *C. albicans* RPS region and the HOK gene is found within the *C. albicans*-specific fingerprinting probe Ca3 that has previously been shown to hybridize very poorly to *C. dubliniensis* genomic DNA (Sullivan *et al.*, 1995).

Table 3. Functional categories of *C. albicans* genes predicted to be of low nucleotide sequence homology or absent in *C. dubliniensis* (normalized ratio <0.25)

Functional category	No. of genes	Percentage of total
Hypothetical genes	134	54.3
Retrotransposon elements	21	8.5
Putative transcriptional regulator	19	7.6
Leucine-rich repeat family (<i>IFA</i>)	18	7.3
GPI-anchored proteins	17	6.9
Transporters	8	3.2
Protein processing/modification	7	2.8
Cell metabolism/biosynthesis	6	2.4
mRNA processing	4	1.6
Cell division	3	1.2
Chromatin/DNA-binding	3	1.2
Morphogenesis related	3	1.2
Cytoskeletal	2	0.8
Mitochondrial	2	0.8
Total	247	100

(i) Putative transcriptional regulators. A group of 19 genes were identified that possessed homology to putative transcriptional regulators. Eight of these regulators had strong homology to genes encoding transcriptional activators in *Saccharomyces cerevisiae* with zinc-finger DNA-binding motifs. A further eight corresponded to a family of genes encoding proteins with homology to a putative *C. albicans* transcriptional activator, *CTA2* (Kaiser *et al.*, 1999). All eight *CTA2*-like genes included for analysis exhibited normalized ratios of <0.25. A CLUSTAL_W-generated alignment of the nucleotide sequences of *CTA21*, *CTA22*, *CTA25* and *CTA26* from *C. albicans* revealed that these sequences were 89% identical. A PCR primer pair (*CTA2F/CTA26R*, Table 2) homologous to conserved sequences in these ORFs did not yield amplicons from *C. dubliniensis* genomic DNA template at a primer annealing temperature of 50°C. A *CTA2* probe was PCR amplified from *C. albicans* template DNA using the *CTA2F/CTA2R* primer pair. This probe which corresponds to >90% of the *CTA2* ORF was at least 90% homologous to six *C. albicans* *CTA2*-like sequences annotated in the CandidaDB. In Southern hybridization experiments this probe revealed multiple hybridizing fragments in *EcoRI*- or *HindIII*-digested *C. albicans* genomic DNA (Fig. 2). Hybridization of the same sequence to *C. dubliniensis* CD36 genomic DNA digested with *EcoRI* or *HindIII* at low stringency did not reveal any hybridizing fragments (Fig. 2). In addition, this *CTA2* probe did not hybridize to genomic DNA from seven additional epidemiologically unrelated *C. dubliniensis* isolates in Southern hybridization experiments (data not shown). These findings are in agreement with the array data that members of this gene family are significantly divergent (i.e. share low nucleotide sequence homology) or are absent in *C. dubliniensis*.

(ii) Putative membrane transporters. Eight genes encoding proteins with strong homology to transporters of nutrients or other small molecules were also found to hybridize poorly with *C. dubliniensis* genomic DNA. These included the genes encoding the oligopeptide transporter *OPT1*, the choline transporters *HNH3* and *HNH4*, the uracil permease *FUR4* and the allantoin permease *DAL52*. The absence of homologous sequences for this group of genes was confirmed in *C. dubliniensis* by Southern hybridization analysis with probes corresponding to the complete *C. albicans* ORF (Fig. 3).

(iii) A leucine-rich repeat family of proteins. A large family of genes encoding proteins with leucine-rich repeats (termed the *IFA* family, Pasteur CandidaDB) in *C. albicans* was also identified. Of 26 *IFA* sequences included for analysis, 18 gave normalized ratios of <0.25 following hybridization of *C. dubliniensis* genomic DNA. Four *IFA* sequences gave intermediate ratios (0.25–0.5) and four sequences (*IFA3*, *IFA13*, *IFA20* and *IFA21*) gave ratios above 0.5. A CLUSTAL_W-generated alignment of four *IFA* sequences with ratios of <0.25 (*IFA1*, *IFA2*, *IFA4* and *IFA5*) revealed that the 3' half of these ORFs exhibited the highest levels of conservation. In order to detect sequences homologous to *IFA* genes in *C. dubliniensis* we amplified a conserved region of 690 bp representing 34% of the *IFA1* ORF using the primer pair *IFA1F/IFA1R* (Table 2). Southern hybridization experiments with this conserved *C. albicans* sequence revealed the presence of a single hybridizing 1.5 kb *EcoRI* fragment in CD36 and in seven additional epidemiologically unrelated *C. dubliniensis* isolates (data not shown). We applied the *IFA1F/IFA1R* primer pair to CD36 template DNA in low-stringency PCR reactions (50°C annealing temperature) and successfully amplified a 728 bp fragment with 70% homology to the *C. albicans* *IFA8* ORF, in keeping with its ratio value between 0.25 and 0.5.

(iv) Genes encoding GPI-anchored proteins. Genes encoding GPI-anchored proteins, including the hypha-specific protein Hyr1 (Bailey *et al.*, 1996), were also identified in our CGH analysis. Two sequences, representing the 3' end and an internal fragment of the *HYR1* gene, were present on the array. Both sequences yielded normalized ratios of <0.3, and no homologous sequence could be identified in *C. dubliniensis* by low-stringency Southern blot analysis of *C. dubliniensis* CD36 genomic DNA with PCR-amplified *C. albicans* *HYR1* gene sequences (nucleotides +95 to +1183, representing 40% of the ORF, chosen because the domain bears most homology to other GPI-anchored protein-encoding genes in *C. albicans*) (Fig. 4a). This was confirmed in Southern hybridization experiments with seven additional epidemiologically unrelated *C. dubliniensis* isolates (data not shown).

Several genes encoding Hyr1-related proteins (termed the *IFF* gene family, Pasteur CandidaDB) were also present on the array. Of the nine *IFF* sequences included for analysis

Table 4. *C. albicans* SC5314 genes predicted to be of low homology (<60% nucleotide sequence identity) or absent in *C. dubliniensis* CD36

Gene identifier*	Assembly 19 orf	Putative or known function†
Unknown function (134)		
IPF3468	orf19.4055	Homology to IPF708
IPF2690.3f/IPF2690.5f	orf19.6614	Contains DEAD helicase box
IPF17640	orf19.3906	Homology to IPF15492
IPF4137.3f	orf19.5370	Homology to ScYBR075w
IPF7945	orf19.6604	NH
IPF7010	orf19.4241	Homology to IPF324.3
IPF17417	orf19.4691	Homology to IPF11506
IPF708	orf19.5370	Homology to ScYBR075w
IPF6387.3f	orf19.584.3	NH
IPF12498.3f/IPF12498.53f	orf19.1677	NH
IPF13810.3	orf19.1825	NH
IPF5661	orf19.7125	NH
IPF2702	orf19.2163	Homology to ScYBR074w
IPF17661	orf19.1348	Homology to ScYBR075w
IPF17272.3f	orf19.3522	NH
IPF13072	orf19.2905	NH
IPF3105	orf19.4439	NH
IPF17488.3f/IPF17488.5f	orf19.1818	Homology to IPF13810
IPF7940	orf19.6608	NH
IPF6266	orf19.6311	Homology to ScYPR036w
IPF11508	orf19.3904	NH
IPF14254	orf19.4768	NH
IPF19766	Not found	NH
IPF11506	orf19.3903	Homology to IPF17417
IPF15772	orf19.4883.3	NH
IPF19377	orf19.3336	NH
IPF10280	orf19.1266	Homology to IPF3748
IPF7804.5f	orf19.3134	NH
IPF19720.3eoc	orf19.4643	NH
IPF4504	orf19.7002	Homology to AnAN3284.2
IPF2815	orf19.6777	NH
IPF17131	orf19.69	NH
IPF9655	orf19.3988	NH
IPF2754	orf19.2585	Homology to ScYER181c
IPF9401	orf19.938	NH
IPF4751	orf19.1999	NH
IPF6325	orf19.1116	NH
IPF13290	orf19.5314	NH
IPF9400	orf19.937	NH
IPF17727	orf19.5190	Part of RPS region
IPF17322.3f	orf19.4068	Homology to IPF13810
IPF2195	orf19.6920	NH
IPF11936.3f	orf19.6277	NH
IPF17794	orf19.6469	NH
IPF2617	orf19.6687	NH
IPF17991	orf19.6465	NH
IPF12093	orf19.4786	NH
IPF14587.3	Not found	NH
IPF20134	orf19.994	NH
IPF11051	orf19.4321	NH
IPF12399	orf19.3210	NH
IPF324.3	orf19.7545	Homology to IPF13810

Table 4. cont.

Gene identifier*	Assembly 19 orf	Putative or known function†
IPF3444.5f	orf19.6184	NH
IPF635	orf19.7614	NH
<i>IFB1</i>	orf19.6703	NH
IPF18833.3f	orf19.1042.1	NH
IPF10231.exon	orf19.426	NH
IPF9211.5f	orf19.3713	NH
IPF15506	orf19.2213	Homology to ScYGR025w
IPF19812	orf19.7304	NH
IPF5373	orf19.5555	NH
IPF13724	orf19.259	Local homology to ScRsa1p
IPF8642	Not found	Homology to IPF10761
IPF243	orf19.3254	NH
IPF3301	orf19.4394	NH
IPF5730	Not found	Homology to ScYNL211c
IPF7578	orf19.4366	Homology to AnAN4487.2
IPF10168.3	orf19.643	NH
IPF14618	orf19.6079	NH
IPF9057	orf19.1189	NH
IPF7539	orf19.4886	NH
IPF5978	orf19.7311	NH
IPF5217	orf19.7270	NH
IPF931	orf19.7567	Homology to ScYDR124w
IPF15255	orf19.4884	Homology to ScYEL007W
IPF2057	orf19.5269	NH
IPF14107	orf19.2449	NH
IPF11756	orf19.7006	NH
IPF15824	orf19.6030	Local homology to ScYKR023w
IPF7338	orf19.3762	Homology to IPF13810
IPF122	orf19.5933	NH
IPF15335	orf19.4970	NH
IPF8741.5f	orf19.4275	NH
IPF1709	orf19.3076	Homology to ScTvpp15 and AnANO175.2
IPF16368.3f	orf19.254	NH
IPF8942	orf19.204	Local homology to ScRim2p
IPF7848	orf19.2558	NH
IPF1742.3f.eoc	orf19.3091	NH
IPF19554.3f	orf19.3371	Homology to AnAN2129.2
IPF13231	orf19.3877	NH
IPF19807	orf19.7032	Homology to ScYLR173
IPF16231	orf19.1913	NH
IPF17542	orf19.5578	NH
IPF2082	orf19.4085	NH
IPF19542.5f	orf19.371	NH
IPF7644	orf19.862	NH
IPF15601	orf19.2433	NH
IPF5453	orf19.5692	Homology to AnAN0905.2
IPF14827	orf19.1611	NH
IPF3733	orf19.5287	NH
IPF11118	orf19.6967	NH
IPF9325	orf19.6288	NH
IPF6070	orf19.4921	Homology to AnAN4487.2
IPF10761	orf19.6021	Homology to IPF8642
IPF7023.3	orf19.134	NH
IPF7201	orf19.2476	NH

Table 4. cont.

Gene identifier*	Assembly 19 orf	Putative or known function†
IPF5248	orf19.7254	NH
IPF14614	orf19.6080.1	NH
<i>IFJ3</i>	orf19.3924	NH
IPF19671	orf19.5854.1	NH
IPF5330	orf19.3448	NH
IPF15540	orf19.258	NH
IPF11347	orf19.4340.1	NH
IPF1548	orf19.951	NH
IPF12829	orf19.721	NH
IPF5217	orf19.7272	NH
IPF7097	orf19.480	Homology to ScYJR056c
IPF15781	orf19.716	NH
IPF4750	orf19.1998	NH
IPF4563.5f	orf19.911	Local homology to <i>ScTOM1</i>
IPF7166	orf19.1202	NH
IPF1717	orf19.3080	Homology to ScYDR180w
IPF2822	orf19.6781	Homology to IPF9894
IPF17064	orf19.6342	NH
IPF6794	orf19.5596	NH
IPF1873	orf19.5902	NH
IPF2856	orf19.3914	Homology to ScYOL139c
IPF9206	orf19.3687	NH
IPF19588	orf19.2318.1	Homology to ScYDR367w
IPF1805	orf19.5003	Homology to ScYGL164c
IPF8420	orf19.929	Contains F-box domain
IPF2195	orf19.6920	NH
IPF171	orf19.3228	Homology to ScYER113c
<i>HOK</i>	orf19.7004	Part of Ca3 fingerprinting probe
Retrotransposon-encoded sequences (21)		
Zorro1b.5f/Zorro1b.3f	orf19.7274	Reverse transcriptase
Zorro2a.3f	orf19.3387	Reverse transcriptase
Zorro2b.5f	orf19.7275	Reverse transcriptase
IPF14825	orf19.1610	Homology to reverse transcriptases
IPF19295.5eoc	orf19.6468	Homology to reverse transcriptases
IPF17652.3	orf19.6078	Homology to reverse transcriptases
<i>Cirt</i>	orf19.4919	Transposase
<i>Cirt1a</i>	orf19.3492	Transposase
<i>Cirt2</i>	orf19.3820	Transposase
<i>Cirt3</i>	orf19.4918	Transposase
<i>Cirt4a</i>	orf19.2839	Transposase
<i>Cirt4b</i>	orf19.2839	Transposase
IPF6235	orf19.5372	<i>C. albicans</i> Tca2 retrotransposon
POL.3	orf19.2372	Pol polyprotein, reverse transcriptase
POL0	orf19.5373	Pol part of pCal retrotransposon
IPF6238	orf19.2374	GAG protein of retrotransposon pCAL
Tca5a	orf19.2427	Polyprotein of Tca5 retrotransposon
IPF2535	Not found	Homology to Tca5 polyprotein (<i>POL</i>) gene
IPF19295.3f	orf19.6469	Homology to Pol polyprotein of <i>A. thaliana</i>
IPF4450	orf19.6078	Pol polyprotein, reverse transcriptase
IPF13885.repeat/IPF13885	orf19.5503	Homology to GAG
Putative transcriptional regulators (19)		
<i>CTA21</i>	orf19.6112	Homology to <i>C. albicans</i> CTA2
<i>CTA22</i>	orf19.3074	Homology to <i>C. albicans</i> CTA2
<i>CTA241.exon1/CTA241.exon2</i>	orf19.5700	Homology to <i>C. albicans</i> CTA2

Table 4. cont.

Gene identifier*	Assembly 19 orf	Putative or known function†
<i>CTA24.3/CTA24</i>	orf19.4054	Homology to <i>C. albicans</i> CTA2
<i>CTA2.5.3f/CTA25</i>	orf19.362	Homology to <i>C. albicans</i> CTA2
<i>CTA26</i>	orf19.7680	Homology to <i>C. albicans</i> CTA2
<i>CTA27</i>	orf19.631	Homology to <i>C. albicans</i> CTA2
<i>CTA29.exon2/CTA29.exon1</i>	orf19.7544	Homology to <i>C. albicans</i> CTA2
<i>RRN3</i>	orf19.1923	ScRRN3, for transcription of rDNA
<i>IPF9315</i>	orf19.4647	Homology to ScHAP3 activator
<i>IPF4708</i>	orf19.5276	Involved in transcriptional elongation
<i>IPF14255</i>	orf19.4767	Zinc-finger protein, GAL4 domain
<i>IPF15920</i>	orf19.4972	Zinc-finger protein
<i>IPF10533.exon1/IPF10533.exon2</i>	orf19.1255	Zinc-finger protein, homology to AnAFLR
<i>IPF16067</i>	orf19.3190	Zinc-finger protein, homology to ScHal9p
<i>IPF9191.3f</i>	orf19.3188	Zinc-finger protein, homology to ScHal9p
<i>IPF8627</i>	orf19.3969	Zinc-finger protein, homology to ScMGA1
<i>IPF19540</i>	orf19.723	Zinc-finger protein
<i>IPF13021</i>	orf19.2647	Zinc-finger protein, homology to ScHAP1
Leucine-rich repeat family (18)		
<i>IFA1</i>	orf19.156	Leucine-rich repeat protein
<i>IFA2</i>	orf19.195	Leucine-rich repeat protein
<i>IFA4</i>	orf19.4510	Leucine-rich repeat protein
<i>IFA5</i>	orf19.6353	Leucine-rich repeat protein
<i>IFA6</i>	orf19.5177	Leucine-rich repeat protein
<i>IFA7</i>	orf19.1326	Leucine-rich repeat protein
<i>IFA9</i>	orf19.2663	Leucine-rich repeat protein
<i>IFA10</i>	orf19.4549	Leucine-rich repeat protein
<i>IFA11</i>	orf19.1596	Leucine-rich repeat protein
<i>IFA12</i>	orf19.5619	Leucine-rich repeat protein
<i>IFA15</i>	orf19.996	Leucine-rich repeat protein
<i>IFA17.5f/IFA17.3f</i>	orf19.4511	Leucine-rich repeat protein
<i>IFA18.3</i>	orf19.4507	Leucine-rich repeat protein
<i>IFA19</i>	orf19.5141	Leucine-rich repeat protein
<i>IFA22</i>	orf19.1002	Leucine-rich repeat protein
<i>IFA24.3</i>	orf19.1596	Leucine-rich repeat protein
<i>IFA25</i>	orf19.5140	Leucine-rich repeat protein
<i>IPF3540</i>	orf19.2814	<i>C. albicans</i> IFA family homologue
GPI-anchored (17)		
<i>PGA15</i>	orf19.2878	Predicted GPI-anchored protein
<i>PGA28</i>	orf19.5144	Predicted GPI-anchored protein
<i>PGA42</i>	orf19.2907	Predicted GPI-anchored protein
<i>PGA44</i>	orf19.1714	Predicted GPI-anchored protein
<i>PGA50</i>	orf19.1824	Predicted GPI-anchored protein
<i>ALS1.3eoc</i>	orf19.5741	GPI-anchored protein, putative adhesin
<i>ALS5</i>	orf19.5736	GPI-anchored protein, putative adhesin
<i>ALS6</i>	orf19.7414	GPI-anchored protein, putative adhesin
<i>ALS7</i>	orf19.7400	GPI-anchored protein, putative adhesin
<i>ALS10</i>	orf19.2355	GPI-anchored protein
<i>ALS11.3f</i>	orf19.5742	GPI-anchored protein
<i>HYR1</i>	orf19.4975	GPI-anchored protein
<i>IFF2</i>	orf19.575	GPI-anchored protein
<i>IFF4</i>	orf19.7472	GPI-anchored protein
<i>IFF7</i>	orf19.3279	GPI-anchored protein
<i>IFF8</i>	orf19.570	GPI-anchored protein
<i>CRH12</i>	orf19.3966	GPI-anchored protein

Table 4. cont.

Gene identifier*	Assembly 19 orf	Putative or known function†
Putative membrane transporters (8)		
<i>OPT1</i>	orf19.2602	Oligopeptide transporter
<i>DAL52</i>	orf19.3208	Putative allantoin permease
IPF11550.3f/IPF11560.5f	orf19.2552	Homology to Ca ²⁺ -transporting ATPases
<i>FUR4</i>	orf19.313	Uracil permease
<i>HN3</i>	orf19.2587	Choline permease
<i>HN4</i>	orf19.2946	Choline permease
IPF1992	orf19.7336	Homology to <i>ScAZR1</i> , drug efflux pump
<i>PTH2</i>	orf19.4231	Proline transport helper
Protein processing/modification (7)		
IPF4710	orf19.5275	Homology to <i>ScVTA1</i>
<i>CTM1</i>	orf19.7612	Homology to <i>ScCTM1</i>
IPF4195	orf19.5441	Homology to <i>ScUlp2p</i>
<i>UBR11.3</i>	orf19.2695	Homology to <i>ScUBR1</i>
<i>PBN1</i>	orf19.3447	Homology to protease B
IPF2997	orf19.6983	Homology to <i>ScReg1p</i> , protein phosphatase
IPF11977	orf19.853	Homology to aspartic proteinases
Cell metabolism/biosynthesis (6)		
IPF19538	orf19.4826	Putative isocitrate dehydrogenase
IPF5239	orf19.7260	Putative aldose reductase
<i>PPX1</i>	orf19.4107	Putative exopolyphosphatase
<i>BIO4</i>	orf19.2590	Dethiobiotin synthetase
IPF4940	orf19.7667	Putative isoamyl acetate esterase
<i>CHS5</i>	orf19.807	Chitin biosynthesis protein
mRNA processing (4)		
IPF4706	orf19.5277	Homology to <i>ScUpf3p</i> ; mRNA decay
IPF1911	orf19.7139	Homology to <i>ScSyf2p</i> ; mRNA splicing
IPF6444	orf19.1900	Homology to <i>ScTgs1p</i>
<i>CUS1</i>	orf19.7581	Spliceosome-associated protein
Chromatin/DNA-binding (3)		
IPF10490	orf19.419	Homology to <i>ScESC1</i>
IPF4805	orf19.3348	Homology to <i>ScNFI1/SIZ1</i>
IPF670	orf19.7598	Homology to <i>ScAHC1</i>
Morphogenesis related (3)		
IPF13247	orf19.3376	Homology to <i>ECE1</i>
IPF13252	orf19.13252	Homology to <i>ECE1</i>
IPF946	orf19.7561	<i>EFG1</i> -dependent transcript <i>EDT1</i>
Cell division (3)		
IPF2971	orf19.4284	<i>Schiz. pombe</i> cyclin c homologue
IPF1760.3f	orf19.4984	Homology to endochitinases
IPF1759.53f	Not found	Homology to endochitinases
Mitochondrial proteins (2)		
IPF5224	orf19.7267	Homology to <i>ScYHR083w</i>
IPF11802	orf19.2797	Homology to <i>ScYDR332w</i>
Cytoskeletal (2)		
IPF11222	orf19.5840	Homology to dynein light-chain proteins
<i>YKE2.3</i>	Not found	Actin-binding protein

*Gene identifiers refer to those in the CandidaDB (<http://genolist.pasteur.fr/CandidaDB/>).

†Gene functions as assigned in the CandidaDB, except where significant homology was independently detected by searches of GenBank or *Saccharomyces* Genome Database (SGD). An indicates homology to *Aspergillus nidulans* genes, Sc indicates homology to *Saccharomyces cerevisiae* genes, NH indicates that no homology was detected.

Fig. 2. Southern hybridization analysis of *C. albicans* and *C. dubliniensis* DNA with a DIG-11-dUTP-labelled probe homologous to nucleotides +1 to +781 of *CTA26*. The blot contains *C. dubliniensis* CD36 genomic DNA and *C. albicans* SC5314 genomic DNA digested with *EcoRI* and *HindIII*, as indicated. Molecular size markers are indicated on the left. Washes were performed at reduced stringency (60 °C in 0.5 × SSC).

(*IFF2* to *IFF11*) only *IFF10.5* and *IFF11* gave ratios above 0.5. Sequence alignments of *C. albicans* *IFF* genes generated with CLUSTAL_W identified *IFF1* as the most likely ancestral gene based on its homology to other members, with the most conserved sequences present in the 5' region. As sequences homologous to this region of *IFF1* were not included on the Eurogentec array, we amplified 844 bp at the 5' end of *IFF1* (~50 % of the ORF) from *C. albicans* SC5314 with the primer pair *IFF1F/IFF1R* (Table 2) and hybridized it to *C. dubliniensis* CD36 genomic DNA to reveal a single hybridizing band (Fig. 4b). We could detect this hybridizing fragment in the genomic DNA of seven additional unrelated *C. dubliniensis* isolates in Southern hybridization experiments. We used the *IFF1F/IFF1R* primer pair in PCR amplification reactions using *C. dubliniensis* CD36 template DNA and successfully generated a 750 bp amplicon with 86 % homology to *C. albicans* *IFF1*.

Other sequences with characterized gene products or functions that could be inferred from homology searches included genes involved in biotin synthesis (*BIO3*, *BIO4*) and several unrelated genes encoding metabolic enzymes (Table 4).

Putative virulence factors

The dataset of genes with normalized ratios of <0.5 was searched to identify genes which had previously been associated with *C. albicans* virulence.

Fig. 3. Southern hybridization analysis of *C. albicans* and *C. dubliniensis* DNA with [α -³²P]dATP-labelled probes corresponding to the complete ORF sequences of the *C. albicans* genes *OPT1*, *FUR4*, *HNM3* and *HNM4*. The ORFs were amplified from *C. albicans* genomic DNA with the primer sets *OPTA/B*, *FUR4A/B*, *HNM3A/B* and *HNM4A/B* (Table 2). Each blot contains *EcoRI*-digested genomic DNA from *C. albicans* SC5314, *C. dubliniensis* CD36 and *C. dubliniensis* CD514. Molecular size markers are indicated on the left. Washes were performed at reduced stringency (60 °C in 0.5 × SSC).

Fig. 4. Southern hybridization analysis of *C. albicans* and *C. dubliniensis* genomic DNA with sequences corresponding to highly conserved regions of *C. albicans* GPI-anchored protein encoding genes. Panel (a) was hybridized with an [α - 32 P]dATP-labelled probe corresponding to nucleotides +95 to +1183 of the *C. albicans* *HYR1* gene. Panel (b) was hybridized with an [α - 32 P]dATP-labelled probe corresponding to nucleotides +1 to +844 of *IFF1*. Each blot contains *C. dubliniensis* CD36 genomic DNA and *C. albicans* SC5314 genomic DNA digested with *EcoRI* and *HindIII*, as indicated. Molecular size markers are indicated on the left. Washes were performed at low stringency (60 °C in 0.5 $\times$ SSC).

(i) Genes encoding putative adhesins. Of the eight sequences with homology to members of the *ALS* gene family of GPI-anchored proteins (encoding putative adhesins) included for analysis, all gave normalized ratios of <0.5, with sequences specific for *ALS1*, *ALS5*, *ALS6* and *ALS7* yielding normalized ratios of <0.25 (Hoyer, 2001). Spots homologous to *ALS2*, *ALS3* and *ALS9* were excluded from our analysis due to poor hybridization with *C. albicans* genomic DNA.

We also investigated whether sequences encoding another group of GPI-anchored proteins were present, namely *HWP1*, which encodes a hyphal adhesin, and related sequences, *RBT1* and *IPF14331* (Braun *et al.*, 2000; Staab *et al.*, 1999). *HWP1* has been associated with virulence in *C. albicans* by mediating adhesion to epithelial cells (Staab *et al.*, 1999). *HWP1* and *RBT1* both yielded normalized ratios of <0.5 (0.48 and 0.37, respectively) in all experiments with *C. dubliniensis* CD36 genomic DNA. In order to identify a homologue of *HWP1*, a primer set designed based on the *C. albicans* *HWP1* sequence (HWP1F/HWP1R, Table 2) was used to amplify a 1.3 kb region of *C. dubliniensis* genomic DNA. The putative 5' upstream region was

also amplified with primers designed based on the sequence of the corresponding *C. albicans* region (APL6F and HWP2, Table 2). An ORF of 1266 bp with homology to *C. albicans* *HWP1* was identified in these sequences. However, the ORF shared only 49% identity with the nucleotide sequence of *C. albicans* *HWP1* due to the presence of several large deletions within the coding sequence (GenBank accession no. AJ632273). The overlapping 5' region amplified from *C. dubliniensis* contained upstream sequences homologous to the *C. albicans* *APL6* gene. This synteny between *HWP1* and *APL6* is conserved in *C. albicans*, and provides further evidence that this gene is a *C. dubliniensis* *HWP1* homologue. However, the predicted protein encoded by the *C. dubliniensis* gene was 421 amino acids in length, 213 residues shorter than the *C. albicans* protein. The first 50 residues of each protein were highly homologous, both containing the KR signature of the KEX2 cleavage site (Fig. 5a). However, the remainder of the N-terminal half of CdHwp1p contained several large deletions compared to the *C. albicans* protein, including most of the region rich in proline, glutamine and aspartate residues (Fig. 5a) (Sundstrom, 2002). Two of these deletions (of 89 bp and 119 bp) spanned the region homologous to the microarray probe and were likely to be responsible for the low signal detected with *C. dubliniensis* genomic DNA. Further deletions were found in the serine-threonine-rich region, although the ω -site for GPI-anchor addition was conserved.

Similarly, a *C. dubliniensis* sequence 513 bp in length was PCR-amplified using the primers RBTF2/RBTR2 designed based on the *C. albicans* *RBT1* sequence (Table 2). This sequence was 45% identical to base pairs +580 to +1410 of the *C. albicans* *RBT1* ORF, the region encoding the internal serine-threonine-rich domain of the *C. albicans* protein. Like the *CdHWP1* gene, this sequence also contained deletions (of 52 bp and 82 bp), when aligned to the homologous *C. albicans* sequence, that were responsible for the poor hybridization of *C. dubliniensis* DNA to the *CaRBT1* array spots.

Southern blot analysis was performed to determine whether single or multiple *HWP1* and *RBT1* homologues could be detected in *C. dubliniensis*. Hybridization of the *C. dubliniensis* *HWP1* amplified sequence to *EcoRI*-digested *C. albicans* genomic DNA revealed a single hybridizing fragment of 4 kb (Fig. 5b). In *C. dubliniensis* genomic DNA, a strongly hybridizing fragment of 9 kb was detected and a second, weakly hybridizing fragment of 5 kb in *EcoRI*-digested DNA (Fig. 5b). This second fragment was identical in size to the fragment detected in Southern blots of *C. dubliniensis* DNA with sequences corresponding to *CaRBT1* (Fig. 5c), indicating that this second hybridizing fragment was likely to correspond to *CdRBT1*.

(ii) Secreted aspartyl proteinases. Sequences homologous to the 10 *C. albicans* secreted aspartyl proteinase (*SAP*)-encoding genes (*SAP1* to *SAP10*) were included on

Fig. 5. (a) Diagram illustrating regions of homology to *C. albicans* CaHwp1p and the extent of deletions in the predicted CdHwp1p protein sequence. The upper rectangular box represents the CaHwp1p protein and shows the position of the KEX2 cleavage site (arrow), the recombinant rHwp1p domain (shaded area) shown to possess transglutaminase substrate activity (Sundstrom, 2002), the serine-threonine-rich region (Ser-Thr rich) and the carboxy-terminal ω -site. The lower boxes represent the homologous regions of the predicted *C. dubliniensis* CdHwp1p protein. The numbers in parentheses indicate the positions of the homologous *C. dubliniensis* protein domains relative to the corresponding *C. albicans* amino acid residues. (b, c) Southern hybridization analysis of *C. albicans* SC5314 and *C. dubliniensis* CD36 genomic DNA with sequences corresponding to (b) *HWP1* and (c) *RBT1*. DNA in (b) was hybridized with an [α - 32 P]dATP-labelled probe corresponding to the entire *C. dubliniensis* *HWP1* ORF. DNA in (c) was hybridized with an [α - 32 P]dATP-labelled probe corresponding to nucleotides +694 to +1410 of *RBT1* amplified from *C. albicans* genomic DNA. Lanes 1 and 2 in both panels contain *Eco*RI-digested genomic DNA from *C. albicans* and *C. dubliniensis*, respectively. Molecular size markers are indicated on the left. Washes were performed at reduced stringency (60 °C in 0.5 $\times$ SSC).

the arrays. All of the *SAP* genes with the exception of *SAP4*, *SAP5* and *SAP6* gave normalized ratios of >0.6 . *SAP5* gave a mean ratio of 0.38 (among genes with intermediate homology) in all *C. dubliniensis* CD36 experiments. We probed the *C. dubliniensis* genome for homologues of *SAP4–6* using PCR primers homologous to conserved regions of these genes (*SAP4*-6F/*SAP4*-6R, Table 2). Amplification using *C. dubliniensis* genomic DNA as the template yielded a PCR product of 750 bp that shared 86% identity with *SAP4* and *SAP6*. Using an inverse PCR strategy (primers InvSAPF/InvSAPR, Table 2) we amplified flanking sequences from *C. dubliniensis* genomic DNA to obtain the complete ORF (GenBank accession no. AJ634382). This *C. dubliniensis* gene was found to lie upstream of the *C. dubliniensis* homologue of the *C. albicans* *SAP1* gene and was equally homologous to *SAP4* and *SAP6* ($\sim 85\%$). This ORF was designated *CdSAP4* as the synteny at this locus with *SAP1* is identical to that at the *SAP4* locus in *C. albicans*. We used the entire *C. dubliniensis* *SAP4* gene as a probe in Southern blots with *C. albicans* and *C. dubliniensis* genomic DNA in order to identify *SAP5* and *SAP6* homologues in *C. dubliniensis*. The *SAP4–6* genes in *C. albicans* are highly homologous (89% nucleotide sequence identity), so we anticipated that the *C. dubliniensis* *SAP4* gene should hybridize strongly to any *C. dubliniensis* *SAP5* or *SAP6* homologues. Indeed, the *C. dubliniensis* *SAP4* gene hybridized to four separate *Kpn*I fragments in *C. albicans*

genomic DNA. Three of these fragments corresponded in length to the predicted sizes of *Kpn*I fragments expected to contain homologues of *SAP4*, *SAP5* and *SAP6* based on the *C. albicans* genome sequences at the Stanford Genome Technology Center. The *SAP4*-hybridizing fragment was less intense than the *SAP5*- and *SAP6*-containing fragments. However, a fourth hybridizing fragment of similar intensity at 7 kb was also detected and could contain a second *SAP4* allele due to an RFLP (Fig. 6). The *C. dubliniensis* *SAP4* gene was hybridized to *C. dubliniensis* genomic DNA digested with *Kpn*I, *Hind*III and several restriction endonucleases that do not cleave within the *CdSAP4* ORF (*Bgl*II, *Spe*I, *Sal*I, *Xba*I); this revealed only one significantly hybridizing band in *C. dubliniensis* genomic DNA (Fig. 6). Furthermore, hybridization of the *C. albicans* *SAP5* and *SAP6* genes to *C. dubliniensis* DNA resulted in hybridization to the same restriction fragment harbouring the *C. dubliniensis* *SAP4* gene (data not shown). These findings were confirmed by Southern blot analysis on eight epidemiologically unrelated isolates of *C. albicans* and *C. dubliniensis* (data not shown).

Hybridization of a second *C. dubliniensis* strain to microarrays

In order to confirm the above data obtained with *C. dubliniensis* CD36 and to investigate the levels of intra-species variation between unrelated *C. dubliniensis* strains, we hybridized genomic DNA from a second *C. dubliniensis*

Fig. 6. Southern hybridization analysis of *C. albicans* and *C. dubliniensis* genomic DNA with sequences corresponding to the *C. dubliniensis* CdSAP4 gene. The blot was hybridized with an [α - 32 P]dATP-labelled probe of the entire CdSAP4 ORF. The blot contains genomic DNA from *C. albicans* SC5314 digested with *KpnI* and genomic DNA from *C. dubliniensis* CD36 digested with *KpnI*, *BglII*, *HindIII*, *SpeI*, *SalI* or *XbaI*, as indicated. The markers on the left side of the panel indicate the predicted positions of the SAP4, SAP5 and SAP6 genes in *C. albicans* SC5314. Molecular size markers are indicated on the right. Washes were performed at reduced stringency (60 °C in 0.5 $\times$ SSC).

isolate, CD514, to these arrays. We chose this strain as it has been shown to be genetically unrelated to *C. dubliniensis* CD36 based on the DNA fingerprint pattern obtained with the *C. dubliniensis* fingerprint probe Cd25. Sheared genomic DNA from *C. dubliniensis* CD514 was labelled with Cy3 and co-hybridized with Cy5-labelled DNA from *C. albicans* SC5314. We compared the dataset from CD514 with that generated from CD36 in order to identify genes unique to each strain. Only three additional genes were discovered that hybridized significantly to CD514 DNA (ratio > 0.59) and that were deemed absent in CD36 (normalized ratio < 0.2, P < 0.035). These genes were IPF4450 and IPF17652.3, with homology to an integrase and a reverse transcriptase, respectively, and the oligopeptide transporter encoding gene *OPT1*. The presence of the *OPT1* sequence in CD514 and its absence in CD36 was confirmed by Southern hybridization with the *C. albicans* *OPT1* sequence (Fig. 2). Conversely, only one sequence, encoding *GIT1* (glycerophosphoinositol transporter), was identified which failed to hybridize with CD514 DNA (ratio 0.112, P value 0.008) and gave significant signals with CD36 DNA (ratio 1.1).

DISCUSSION

Phylogenetic analysis of rRNA sequences has confirmed that *C. dubliniensis* and *C. albicans* are the two most closely related *Candida* species of clinical importance in humans

(Sullivan *et al.*, 1995, 2004). However, whereas *C. albicans* is the most significant yeast pathogen, responsible for superficial and deep-seated infections, *C. dubliniensis* is of lesser clinical importance in mucosal infections in non-HIV-infected patients, and is relatively insignificant in the case of bloodstream infection (Kibbler *et al.*, 2003; Meis *et al.*, 2000). This apparent lower virulence of *C. dubliniensis* is also evident in data from animal model infection studies. However, the exact reasons why *C. albicans* is more virulent are not clear. In this study we have utilized recently available *C. albicans* whole-genome DNA microarrays to investigate and identify genomic differences between *C. albicans* and *C. dubliniensis* that could account, at least in part, for the enhanced virulence potential of *C. albicans* relative to *C. dubliniensis* and for the differences in epidemiology between the two species.

The findings presented in this study obtained by CGH reinforce the phylogenetic data that originally implied the close relatedness of the two organisms (Gilfillan *et al.*, 1998; Sullivan *et al.*, 1995). Our data show that only 4.4 % of *C. albicans* sequences analysed in our studies (normalized ratio < 0.25) were likely to be absent or highly divergent (< 60 % homologous at the nucleotide sequence level) in *C. dubliniensis*. That the vast majority of *C. albicans* genes are highly conserved in *C. dubliniensis* indicates that the two species have probably only diverged relatively recently and thus are likely to inhabit similar environments in the human body. Thus only a small subset of *C. albicans* genes seem to be unique to this species and could possibly be important contributory factors to the greater success of *C. albicans* as a commensal on the human mucosal epithelium and as a pathogen in compromised hosts. Surprisingly, some spots exhibited reproducibly higher fluorescent signals with *C. dubliniensis* genomic DNA. There are several possible reasons for these findings, including cross-hybridization of paralogous *C. dubliniensis* sequences to the same spot, increases in gene copy number or differences in ploidy.

Of the 247 *C. albicans* genes identified that were predicted to have less than 60 % homology at the nucleotide sequence level or even possibly be absent in *C. dubliniensis*, 134 were hypothetical genes of unknown function (Table 3). However, 43 of these hypothetical ORFs were conserved, with homology to genes in *S. cerevisiae*, *Aspergillus nidulans* or paralogous sequences in the *C. albicans* genome. Of the 113 genes with a confirmed or hypothetical function, few were identified that corresponded to housekeeping genes involved in central metabolism, cell structure, or molecular biosynthesis. However, several transporter-encoding genes involved in nutrient uptake were identified as being absent, including two of the four genes encoding choline permeases in *C. albicans* (*HNM3*, *HNM4*), a uracil permease (*FUR4*) and an allantoin permease (*DAL52*). The only confirmed intrastrain difference between the two Cd25 fingerprint group *C. dubliniensis* strains analysed here was the presence of sequences homologous to the

oligopeptide transporter *OPT1* (Lubkowitz *et al.*, 1997) in the Cd25 fingerprint group II strain CD514 which were absent in the fingerprint group I isolate CD36. It is not known whether the absence of these genes could affect the ability of *C. dubliniensis* to grow relative to *C. albicans* *in vivo*, since, for example, our data suggest other genes encoding choline (*HN2*) and allantoin permeases (*DAL51*) are likely to be present in *C. dubliniensis*. *C. dubliniensis* also seems to lack sequences involved in the biosynthesis of biotin (*BIO3* encoding DAPA aminotransferase and *BIO4* encoding dethiobiotin synthetase). Although biotin is required for growth, *C. albicans* and *C. dubliniensis* probably acquire sufficient biotin from exogenous sources in the oral cavity, most likely from commensal bacteria (Phalip *et al.*, 1999).

Twenty-one sequences corresponded to genes present in retrotransposons of *C. dubliniensis*, indicating that since their divergence the genomes of the two species may have acquired different mobile genetic elements.

Seventeen sequences homologous to genes encoding various GPI-anchored proteins were identified in our analysis as being absent or of low homology in *C. dubliniensis* by a combination of array hybridization data, PCR analysis and Southern blot analysis (Sundstrom, 2002). Poor *C. dubliniensis* hybridization signals were detected from sequences homologous to the *C. albicans* hyphal specific *HYR1* gene (mean ratio 0.13), and no homologous gene was identified in *C. dubliniensis* following Southern hybridization with conserved *C. albicans* *HYR1* sequences (Bailey *et al.*, 1996). Sequences corresponding to several *HYR1*-related GPI-anchored proteins in *C. albicans* also exhibited poor hybridization signals with *C. dubliniensis* genomic DNA (*IFF* family genes). Although specific functions have not been assigned to proteins encoded by these genes, their likely location on the cell surface indicates possible roles in maintaining cell wall integrity, environmental signalling or adhesion to host surfaces.

Interestingly, subsequent analysis (Southern blotting) with sequences homologous to the *C. albicans* *IFF1* gene (absent from Eurogentec microarrays) identified a homologous gene in *C. dubliniensis* for which sequences were later identified by PCR. These data suggest that at least one *IFF*-like gene is present in the *C. dubliniensis* genome. This may represent an ancestral *IFF*-related gene; however, additional *IFF*-related genes may be present in the *C. dubliniensis* genome but may be difficult to detect by CGH as they could have diverged more extensively than essential housekeeping genes with greater sequence-based constraints on protein function. A similar conclusion could be reached with regard to sequences homologous to genes encoding proteins of the α -agglutinin-like ALS family of adhesins. The ALS probes on the Eurogentec arrays used in this study consist of sequences from the 3' region of the ORFs, which within the ALS family are the least conserved regions (Hoyer *et al.*, 2001). By Southern

hybridization analysis Hoyer and co-workers noted that the 3' regions of the *C. albicans* ALS genes are poorly conserved in *C. dubliniensis*. We observed low hybridization ratios (<0.25) for several members of this family including *ALS1*, *ALS5*, *ALS6* and *ALS7*. However, Hoyer and co-workers identified partial 5' nucleotide sequences for three ALS homologues in *C. dubliniensis* (*ALSD1*, *ALSD2* and *ALSD3*). Their study demonstrated that *ALSD1* is closely related to *ALS6* and *ALSD3* is closely related to *ALS4*. The present study confirms their findings that the 3' regions of the ALS genes are poorly conserved in *C. dubliniensis*, but does not provide further evidence for the existence of other *C. dubliniensis* ALS homologues. Since the microarray DNA spots correspond to 300–400 bp regions of each gene, our data reflect differences present in these regions only.

As the CGH data obtained for the ALS gene family demonstrate, data indicating the absence or divergence of a particular gene require confirmation as these regions may encompass non-conserved regions of the gene. Conversely, there may be divergent regions in many genes that remain undetected as they lie outside the regions compared here. Similarly, minor genetic differences (e.g. point mutations) and differences in non-translated regions that cannot be detected using these methods could also influence virulence and epidemiology. In addition, phenotype can also be influenced by post-transcriptional events unrelated to DNA sequence.

Low hybridization ratios were also observed for the *HWP1* gene and the related sequences *RBT1* and *IPF14331* (Braun & Johnson, 1997; Staab *et al.*, 1999; Sundstrom, 2002). The *HWP1*-encoded protein has been identified as a hypha-specific substrate for host transglutaminases involved in covalent adhesion to host cells. However, the functions of the other two gene products are as yet uncharacterized. In this study we identified the *C. dubliniensis* *HWP1* homologue. The *C. dubliniensis* *HWP1* gene hybridized poorly to the *C. albicans* array *HWP1* sequences due to the presence of large deletions in the *C. dubliniensis* ORF. The predicted translated protein encoded by *CdHWP1* contains several large deletions compared to the *C. albicans* protein (Fig. 5a). These deletions lie within the N-terminal glutamine- and proline-rich repeat domain containing the transglutaminase substrate activity and the internal serine- and threonine-rich domain. It will be of interest to determine whether the transglutaminase substrate activity of the *C. dubliniensis* homologue is affected by the presence of deletions in glutamine-rich regions of the N-terminus. A defect in the ability of *C. dubliniensis* to form stable attachments to oral epithelium may partly explain its reduced prevalence in the oral cavities of healthy individuals and patients with oral disease.

One of the most intensely studied virulence attributes of *C. albicans* is the ability to secrete aspartyl proteinases, encoded by 10 separate genes (Naglik *et al.*, 2003). Sequences from all 10 *SAP* genes were present on the array.

Sequences from only one of these genes, *SAP5*, gave consistently low signals from *C. dubliniensis* hybridizing DNA. *SAP5* is a member of the *SAP4–6* subfamily of proteinases, which are all highly homologous at the nucleotide sequence level and preferentially expressed by hyphae (Hube *et al.*, 1994). In our efforts to determine if *SAP5* was present in *C. dubliniensis*, we identified a gene that is most homologous to *SAP4* and *SAP6*, which we designated *CdSAP4* as the ORF was located upstream of *CdSAP1*, identical to the synteny observed in *C. albicans* (Miyasaki *et al.*, 1994). Southern hybridization analysis with this *CdSAP4* sequence revealed that it could hybridize to multiple fragments of *C. albicans* restriction-endonuclease-digested DNA corresponding to sequences of *SAP4*, *SAP5* and *SAP6*. Such cross-hybridization is likely to be responsible for the strong signal detected from spots representing *SAP6* on the *C. albicans* microarray. However, the *CdSAP4* sequence consistently hybridized to only one single band (between ~2 and 10 kb) in Southern hybridization experiments with *C. dubliniensis* genomic DNA. These data indicate that only one gene with strong homology to the *SAP4–6* subfamily exists in *C. dubliniensis*. Attempts to identify the corresponding genomic loci of putative *SAP5* and *SAP6* genes in *C. dubliniensis* by low-stringency PCR were unsuccessful (data not shown).

Together, the *SAP4–6* subfamily has been shown to play an important role in the establishment of *C. albicans* systemic infections in mice and *SAP6* has been shown to be the most important gene within this family in the establishment of murine intraperitoneal infection (Felk *et al.*, 2002; Sanglard *et al.*, 1997). As *C. dubliniensis* only possesses one gene with homology to *SAP4–6*, this could partly explain why *C. dubliniensis* is less able than *C. albicans* to establish systemic infection. *In vivo* virulence studies and epidemiological data support this hypothesis, since *C. dubliniensis* is less virulent than *C. albicans* in a murine systemic model of infection and the incidence of recovery of this organism from human blood cultures is extremely low (Gilfillan *et al.*, 1998; Kibbler *et al.*, 2003). The absence of these hypha-specific proteinases in *C. dubliniensis* may affect the ability of its hyphae to penetrate host tissues, acquire nutrients or evade killing by macrophages. We are currently testing the role of *C. dubliniensis* Sap proteins in infection models.

Differences in gene regulation have not been explored in any great detail in *C. dubliniensis* to date. The array CGH data indicates the presence of genes homologous to many of the transcription factors involved in regulating hypha formation in *C. albicans* (e.g. *EFG1*, *CPH1*, *TUP1*). However, poor hybridization signals were detected from several other genes encoding putative transcriptional regulators that could affect regulatory circuits in *C. dubliniensis*. Eight of these sequences had homology to genes encoding proteins with zinc-finger DNA-binding motifs in *S. cerevisiae*. Another group of genes with homology to the putative *C. albicans* transcriptional activator encoding

gene *CTA2* were also identified. *CTA2* (GenBank ID AJ006637) was identified by Kaiser *et al.* (1999) in a one-hybrid screen in *S. cerevisiae* for *C. albicans* proteins with transcriptional activating properties. A family of possibly up to 10 genes with strong homology to *CTA2* has been identified in *C. albicans*. Twelve sequences homologous to these genes were included in our analysis and all gave normalized ratios of <0.25. Southern hybridization also failed to identify any sequences with significant homology to these genes in *C. dubliniensis*. Although the function of these proteins has yet to be confirmed, the absence or divergence of a large family of transcriptional activators in *C. dubliniensis* could have major implications for the growth and virulence of this fungus.

In the present study, *C. albicans* DNA microarrays facilitated a whole-genome comparison between *C. albicans* and its close relative *C. dubliniensis*. Our experiments have revealed the absence and divergence of several genes and gene families in *C. dubliniensis*. These include putative virulence factors and many genes specific for or preferentially expressed in the hyphal phase, such as *SAP5*, *SAP6*, *HWPI* and *HYR1*. *C. dubliniensis* is generally less efficient than *C. albicans* at forming hyphae in response to serum and the absence of these hypha-regulated genes may also indicate that *C. dubliniensis* hyphae are less specialized as virulence-promoting structures (Gilfillan *et al.*, 1998). We have endeavoured to confirm the absence or divergence of genes directly involved in virulence (e.g. *HWPI*, *SAP5*); however, conclusive confirmation of this data will have to await the completion of the *C. dubliniensis* genome sequencing project which is currently under way (http://www.sanger.ac.uk/Projects/C_dubliniensis/). When complete this project will facilitate more thorough comparative genomic analysis and will allow the identification of *C. dubliniensis*-specific genes that could be responsible for phenotypic differences. Until then, this dataset represents a framework for further investigations into genetic and phenotypic differences between *C. albicans* and *C. dubliniensis*.

ACKNOWLEDGEMENTS

This study was supported by the Microbiology Research Unit, Dublin Dental School and Hospital.

REFERENCES

- Al Mosaid, A., Sullivan, D., Salkin, I. F., Shanley, D. & Coleman, D. C. (2001). Differentiation of *Candida dubliniensis* from *Candida albicans* on Staib agar and caffeic acid-ferric citrate agar. *J Clin Microbiol* **39**, 323–327.
- Al Mosaid, A., Sullivan, D. J. & Coleman, D. C. (2003). Differentiation of *Candida dubliniensis* from *Candida albicans* on Pal's agar. *J Clin Microbiol* **41**, 4787–4789.
- Alves, S. H., Milan, E. P., de Laet Sant'Ana, P., Oliveira, L. O., Santurio, J. M. & Colombo, A. L. (2002). Hypertonic Sabouraud broth as a simple and powerful test for *Candida dubliniensis* screening. *Diagn Microbiol Infect Dis* **43**, 85–86.

- Bailey, D. A., Feldman, P. J. F., Bovey, M., Gow, N. A. R. & Brown, A. J. P. (1996). The *Candida albicans* HYR1 gene, which is activated in response to hyphal development, belongs to a gene family encoding yeast cell wall proteins. *J Bacteriol* **178**, 5353–5360.
- Braun, B. R. & Johnson, A. D. (1997). Control of filament formation in *Candida albicans* by the transcriptional repressor *TUP1*. *Science* **277**, 105–109.
- Braun, B. R., Head, W. S., Wang, M. X. & Johnson, A. D. (2000). Identification and characterization of *TUP1*-regulated genes in *Candida albicans*. *Genetics* **156**, 31–44.
- Coleman, D. C., Sullivan, D. J., Bennett, D. E., Moran, G. P., Barry, H. J. & Shanley, D. B. (1997). Candidiasis: the emergence of a novel species, *Candida dubliniensis*. *AIDS* **11**, 557–567.
- Daran-Lapujade, P., Daran, J. M., Kotter, P., Petit, T., Piper, M. D. & Pronk, J. T. (2003). Comparative genotyping of the *Saccharomyces cerevisiae* laboratory strains S288C and CEN.PK113-7D using oligonucleotide microarrays. *FEMS Yeast Res* **4**, 259–269.
- Dong, Y., Glasner, J. D., Blattner, F. R. & Triplett, E. W. (2001). Genomic interspecies microarray hybridization: rapid discovery of three thousand genes in the maize endophyte *Klebsiella pneumoniae* 342, by microarray hybridization with *Escherichia coli* K-12 open reading frames. *Appl Environ Microbiol* **67**, 1911–1921.
- Donnelly, S. A., Sullivan, D. J., Shanley, D. B. & Coleman, D. C. (1999). Phylogenetic analysis and rapid identification of *Candida dubliniensis* based on analysis of *ACT1* intron and exon sequences. *Microbiology* **145**, 1871–1882.
- Felk, A., Kretschmar, M., Albrecht, A. & 7 other authors (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.
- Gallagher, P. J., Bennett, D. E., Henman, M. C., Russell, R. J., Flint, S. R., Shanley, D. B. & Coleman, D. C. (1992). Reduced azole susceptibility of *Candida albicans* from HIV-positive patients and a derivative exhibiting colony morphology variation. *J Gen Microbiol* **138**, 1901–1911.
- Gee, S. F., Joly, S., Soll, D. R., Meis, J. F., Verweij, P. E., Polacheck, I., Sullivan, D. J. & Coleman, D. C. (2002). Identification of four distinct genotypes of *Candida dubliniensis* and detection of microevolution *in vitro* and *in vivo*. *J Clin Microbiol* **40**, 556–574.
- Gilfillan, G. D., Sullivan, D. J., Haynes, K., Parkinson, T., Coleman, D. C. & Gow, N. A. R. (1998). *Candida dubliniensis*: phylogeny and putative virulence factors. *Microbiology* **144**, 829–838.
- Hannula, J., Saarela, M., Dogan, B., Paatsama, J., Koukila-Kahkola, P., Pirinen, S., Alakomi, H. L., Perheentupa, J. & Asikainen, S. (2000). Comparison of virulence factors of oral *Candida dubliniensis* and *Candida albicans* isolates in healthy people and patients with chronic candidosis. *Oral Microbiol Immunol* **15**, 238–244.
- Higgins, D. G. & Sharp, P. M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237–244.
- Hoyer, L. L. (2001). The *ALS* gene family of *Candida albicans*. *Trends Microbiol* **9**, 176–180.
- Hoyer, L. L., Fundyga, R., Hecht, J. E., Kapteyn, J. C., Klis, F. M. & 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, D. A., Brown, A. J. P. & Gow, N. A. R. (1994). Expression of seven members of the gene family encoding secretory aspartyl proteinases in *Candida albicans*. *Mol Microbiol* **14**, 87–99.
- Jabra-Rizk, M. A., Ferreira, S. M., Sabet, M., Falkler, W. A., Merz, W. G. & Meiller, T. F. (2001). Recovery of *Candida dubliniensis* and other yeasts from human immunodeficiency virus-associated periodontal lesions. *J Clin Microbiol* **39**, 4520–4522.
- Kaiser, B., Munder, T., Saluz, H.-P., Kunkel, W. & Eck, R. (1999). Identification of a gene encoding the pyruvate decarboxylase gene regulator CaPdc2p from *Candida albicans*. *Yeast* **15**, 585–591.
- Kibbler, C. C., Seaton, S., Barnes, R. A., Gransden, W. R., Holliman, R. E., Johnson, E. M., Perry, J. D., Sullivan, D. J. & Wilson, J. A. (2003). Management and outcome of bloodstream infections due to *Candida* species in England and Wales. *J Hosp Infect* **54**, 18–24.
- Kurzai, O., Heinz, W. J., Sullivan, D. J., Coleman, D. C., Frosch, M. & Muhlischlegel, F. A. (1999). Rapid PCR test for discriminating between *Candida albicans* and *Candida dubliniensis* isolates using primers derived from the pH-regulated *PHR1* and *PHR2* genes of *C. albicans*. *J Clin Microbiol* **37**, 1587–1590.
- Lubkowitz, M. A., Hauser, L., Breslav, M., Naider, F. & Becker, J. M. (1997). An oligopeptide transporter gene from *Candida albicans*. *Microbiology* **143**, 387–396.
- Meis, J. F. G. M., Lunel, F. M. V., Verweij, P. E. & Voss, A. (2000). One-year prevalence of *Candida dubliniensis* in a Dutch university hospital. *J Clin Microbiol* **38**, 3139–3140.
- Miyasaki, S. H., White, T. C. & 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.
- Moran, G. P., Sanglard, D., Donnelly, S. M., Shanley, D. B., Sullivan, D. J. & Coleman, D. C. (1998). Identification and expression of multidrug transporters responsible for fluconazole resistance in *Candida dubliniensis*. *Antimicrob Agents Chemother* **42**, 1819–1830.
- Moran, G., Sullivan, D., Morrischauser, J. & Coleman, D. (2002). The *Candida dubliniensis* *CdCDR1* gene is not essential for fluconazole resistance. *Antimicrob Agents Chemother* **46**, 2829–2841.
- Murray, A. E., Lies, D., Li, G., Nealson, K., Zhou, J. & Tiedje, J. M. (2001). DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. *Proc Natl Acad Sci U S A* **98**, 9853–9858.
- Naglik, J. R., Challacombe, S. J. & Hube, B. (2003). *Candida albicans* secreted aspartyl proteinases in virulence and pathogenesis. *Microbiol Mol Biol Rev* **67**, 400–428.
- Perea, S., Lopez-Ribot, J. L., Wickes, B. L., Kirkpatrick, W. R., Dib, O. P., Bachmann, S. P., Keller, S. M., Martinez, M. & Patterson, T. F. (2002). Molecular mechanisms of fluconazole resistance in *Candida dubliniensis* isolates from human immunodeficiency virus-infected patients with oropharyngeal candidiasis. *Antimicrob Agents Chemother* **46**, 1695–1703.
- Pfaller, M. A. & Diekema, D. J. (2004). Twelve years of fluconazole in clinical practice: global trends in species distribution and fluconazole susceptibility of bloodstream isolates of *Candida*. *Clin Microbiol Infect* **10** Suppl 1, 11–23.
- Phalip, V., Kuhn, I., Lemoine, Y. & Jeltsch, J.-M. (1999). Characterization of the biotin biosynthesis pathway in *Saccharomyces cerevisiae* and evidence for a cluster containing *BIO5*, a novel gene involved in vitamer uptake. *Gene* **232**, 43–51.
- Pinjon, E., Sullivan, D., Salkin, I., Shanley, D. & Coleman, D. (1998). Simple, inexpensive, reliable method for differentiation of *Candida dubliniensis* from *Candida albicans*. *J Clin Microbiol* **36**, 2093–2095.
- Pinjon, E., Moran, G. P., Jackson, C. J., Kelly, S. L., Sanglard, D., Coleman, D. C. & Sullivan, D. J. (2003). Molecular mechanisms of itraconazole resistance in *Candida dubliniensis*. *Antimicrob Agents Chemother* **47**, 2424–2437.

- Sanglard, D., Hube, B., Monod, M., Odds, F. C. & Gow, N. A. R. (1997).** A triple deletion of the secreted aspartyl proteinase genes *SAP4*, *SAP5* and *SAP6* of *Candida albicans* causes attenuated virulence. *Infect Immun* **65**, 3539–3546.
- Sebti, A., Kiehn, T. E., Perlin, D., Chaturvedi, V., Wong, M., Doney, A., Park, S. & Sepkowitz, K. A. (2001).** *Candida dubliniensis* at a cancer center. *Clin Infect Dis* **32**, 1034–1038.
- Soll, D. R. (2000).** The ins and outs of DNA fingerprinting the infectious fungi. *Clin Microbiol Rev* **13**, 332–370.
- Southern, E. (1975).** Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* **98**, 503–517.
- Staab, J. F., Bradway, S. D., Fidel, P. L. & Sundstrom, P. (1999).** Adhesive and mammalian transglutaminase substrate properties of *Candida albicans* HWP1. *Science* **283**, 1535–1538.
- Staib, P., Moran, G. P., Sullivan, D. J., Coleman, D. C. & Morschhauser, J. (2001).** Isogenic strain construction and gene targeting in *Candida dubliniensis*. *J Bacteriol* **183**, 2859–2865.
- Sullivan, D. J., Westerneng, T. J., Haynes, K. A., Bennett, D. E. & Coleman, D. C. (1995).** *Candida dubliniensis* sp. nov.: phenotypic and molecular characterization of a novel species associated with oral candidosis in HIV-infected individuals. *Microbiology* **141**, 1507–1521.
- Sullivan, D. J., Moran, G. P., Pinjon, E., Al-Mosaid, A., Stokes, C., Vaughan, C. & Coleman, D. C. (2004).** Comparison of epidemiology, drug resistance mechanisms, and virulence of *Candida dubliniensis* and *Candida albicans*. *FEMS Yeast Res* **4**, 369–376.
- Sundstrom, P. (2002).** Adhesion in *Candida* spp. *Cell Microbiol* **4**, 461–469.
- Vilela, M. M., Kamei, K., Sano, A., Tanaka, R., Uno, J., Takahashi, I., Ito, J., Yarita, K. & Miyaji, M. (2002).** Pathogenicity and virulence of *Candida dubliniensis*: comparison with *C. albicans*. *Med Mycol* **40**, 249–257.
- Willis, A. M., Coulter, W. A., Sullivan, D. J., Coleman, D. C., Hayes, J. R., Bell, P. M. & Lamey, P. J. (2000).** Isolation of *C. dubliniensis* from insulin-using diabetes mellitus patients. *J Oral Pathol Med* **29**, 86–90.