# CandidaDB: a multi-genome database for Candida species and related Saccharomycotina

Tristan Rossignol<sup>1</sup>, Pierre Lechat<sup>2</sup>, Christina Cuomo<sup>3</sup>, Qiandong Zeng<sup>3</sup>, Ivan Moszer<sup>2</sup> and Christophe d'Enfert<sup>1,\*</sup>

<sup>1</sup>Unité Biologie et Pathogénicité Fongiques, INRA USC2019, Institut Pasteur, Paris, France, <sup>2</sup>Plate-forme Intégration et Analyse Génomiques, Pasteur Génopole IIe-de-France, Institut Pasteur, Paris, France and <sup>3</sup>Broad Institute of MIT and Harvard, Cambridge, Massachussets, United States of America

Received September 11, 2007; Revised October 22, 2007; Accepted October 24, 2007

# ABSTRACT

CandidaDB (http://genodb.pasteur.fr/CandidaDB) was established in 2002 to provide the first genomic database for the human fungal pathogen Candida albicans. The availability of an increasing number of fully or partially completed genome sequences of related fungal species has opened the path for comparative genomics and prompted us to migrate CandidaDB into a multi-genome database. The new version of CandidaDB houses the latest versions of the genomes of C. albicans strains SC5314 and WO-1 along with six genome sequences from species closely related to C. albicans that all belong to the CTG clade of Saccharomycotina-Candida tropicalis, Candida (Clavispora) lusitaniae, Candida (Pichia) guillermondii, Lodderomyces elongisporus, Debaryomyces hansenii, Pichia stipitis-and the reference Saccharomyces cerevisiae genome. CandidaDB includes sequences coding for 54170 proteins with annotations collected from other databases, enriched with illustrations of structural features and functional domains and data of comparative analyses. In order to take advantage of the integration of multiple genomes in a unique database, new tools using pre-calculated or userdefined comparisons have been implemented that allow rapid access to comparative analysis at the genomic scale.

# INTRODUCTION

*Candida* species are the most important opportunistic fungal pathogens of humans responsible for superficial and systemic infections (1). Among these species, *Candida albicans* is responsible for the majority of infections, but other species are becoming increasingly common (1). Because of its predominance, *C. albicans* has been the

focus of genomic and molecular studies over the last 20 years, becoming a model organism for other pathogenic *Candida* species and fungal pathogens. The *C. albicans* genome was made publicly available by the Stanford Genome Technology Center at the end of the 1990s and different assemblies and annotations have been released since (2–4). This has been accompanied by the implementation of two main genomic databases: CandidaDB (5) and the Candida Genome Database (6,7).

As infections due to non-*albicans Candida* in hospitals have increased (8), research on these emerging species has recently developed. Genome sequencing projects for these species, as well as related non-pathogenic yeast species, have been completed or are nearing completion (4,9-12). The availability of numerous related genomes paves the way for comparative genomic approaches that have already contributed to our understanding of the evolutionary processes that underlie speciation in the *Sachharomycotina* subphylum (10,13-15). Applied to closely-related pathogenic and non-pathogenic yeast species, comparative genomics should provide insights in virulence processes.

To date, most yeast genomes are available at different databases and there is no resource that enables online comparative analysis. The current aim of the CandidaDB database is to provide such a comparative resource for species of the CTG clade of the subphylum Saccharo*mycotina* that is characterized by the translation of the CUG codon into serine instead of leucine. The CTG clade includes C. albicans and several of the most important human pathogenic fungi (16-18). CandidaDB provides genome sequences of four pathogenic [C. albicans, Candida tropicalis, Candida (Clavispora) lusitaniae, Candida (Pichia) guillermondii] and three non-pathogenic (Lodderomyces elongisporus, Debaryomyces hansenii, Pichia stipitis) species belonging to the CTG clade (Table 1). It also provides the Saccharomyces cerevisiae genome sequence as a reference (19). CandidaDB includes sequences coding for 54170 proteins with annotations collected from other databases. It has been enriched with

\*To whom correspondence should be addressed. Tel: +33 (0)1 40 61 32 57; Fax: +33 (0)1 45 68 89 38; Email: denfert@pasteur.fr

© 2007 The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/ by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Species	Strain	Number of proteins	Number of chromosomes and/or supercontigs	Status and release date	Sequencing center/Database repository	Database links
Candida albicans	SC5314	6098	8	Draft assembly 13 September 2006	CGD	http://www.candidagenome.org/
Candida albicans	WO1	6159	16	Draft assembly 15 March 2006	Broad Institute	http://www.broad.mit.edu/annotation/ genome/candida_albicans/
Candida guilliermondii	ATCC6260	5920	9	Draft assembly 15 March 2006	Broad Institute	http://www.broad.mit.edu/annotation/ genome/candida_guilliermondii/
Candida tropicalis	MYA-3404	6258	23	Draft assembly 12 June 2006	Broad Institute	http://www.broad.mit.edu/annotation/ genome/candida_tropicalis/
Candida lusitaniae	ATCC42720	5941	9	Draft assembly 25 January. 2006	Broad Institute	http://www.broad.mit.edu/annotation/ genome/candida_lusitaniae/
Debaryomyces hansenii	CBS767	6318	7	Complete 3 July 2004	Génolevures	http://cbi.labri.fr/Genolevures/elt/DEHA
Pichia stipitis	CBS 6054	5816	9	Complete 17 April 2007	JGI	http://genome.jgi-psf.org/Picst3/ Picst3.home.html
Lodderomyces elongisporus	NRLL YB-4239	5802	27	Draft assembly 12 June 2006	Broad Institute	http://www.broad.mit.edu/annotation/ genome/lodderomyces_elongisporus/
Saccharomyces cerevisiae	S288C	5858	16	Complete 27 March 2007	SGD	http://www.yeastgenome.org/
Total	9	54 170	124			

Table 1. Characteristics of the nine genomes available in the current release of CandidaDB

illustrations of structural features and functional domains and tools for sequence comparisons and analysis. Moreover, new tools for comparative genomics have been implemented in order to take advantage of the integration of multiple genomes in a unique database. Importantly, pre-calculated comparisons provide rapid access to comparative analysis at the protein and genomic scale.

# SOURCE DATA AND COMPATIBILITY WITH OTHER DATABASES

Eight publicly available genome sequences of seven closely related species belonging to the CTG clade are included in the new release of CandidaDB: the genomes of C. albicans strains SC5314 (2) and WO1 (20); three genomes of other pathogenic species, C. tropicalis strain MYA-3404 (21), C. lusitaniae strain ATCC42720 (22) and C. guilliermondii strain ATCC6260 (23); and the genomes of three non-pathogenic species, L. elongisporus strain NRLL YB-4239 (24), an ascososporogenous species, D. hansenii strain CBS767 (10), a halotolerant yeast found in fish and salted dairy products that have a role in agro-food processes and Pichia stipitis strain CBS6054 (12), a xylose fermenting yeast. The new release of CandidaDB also includes the S. cerevisiae strain S288C genome (19) in order to take advantage of the high level of annotation provided for this species that is not part of the CTG clade but is part of the Saccharomycotina subphylum (17). These genome sequences and associated annotations were obtained from the sources indicated in Table 1 that

summarizes the general information for the nine genomes available in the current version of CandidaDB.

The new version of CandidaDB uses Assembly 20 of the genome sequence of *C. albicans* strain SC5314 genome available at the Candida Genome Database (CGD) (4,7). While previous releases of CandidaDB used annotations contributed by the Galar Fungail consortium (5), CandidaDB now uses sequences, descriptions, accession numbers and annotations available at CGD which is the reference depository site for *C. albicans*. This allows homogenization of the nomenclature for this organism and will simplify literature curation. Accession numbers of previous CandidaDB releases are still available as synonyms.

The genomes of *P. stipitis*, *D. hansenii* and *S. cerevisiae* available through CandidaDB are considered completed and have been published (10,12,19), while the other genomes are draft assemblies, close to completion and with a low number of contigs. CandidaDB aims to follow the usual accession number for Open Reading Frames (ORFs) provided by the institutions which performed the sequences, for better clarity, inter-database relations and faster update procedures.

#### **IMPLEMENTATION**

CandidaDB is based on the general data frame called GenoList (25). GenoList is an integrated environment for multiple genomes based on a relational database run through a web user interface that provides comparative genomic and proteomic tools in complement to the gene

descriptions. Structure and design are detailed in the accompanying paper (25). GenoList has been originally developed as a multigenome database for comparative analysis of bacterial genomes (25) and has been adapted to eukaryotes in order to manage the CandidaDB database.

When connecting to CandidaDB, users are prompted to register and provide a login and password. Although this is optional and no tracking of the registered users is performed, it allows users to specify parameters for CandidaDB usage (see subsequently) and maintain these parameters upon return to the database. Upon registered or unregistered login, users have access to a web interface that is composed of a main window allowing different forms of queries and analysis at the gene, genome and multi-genome scale. Results of the queries are presented in the main window as gene lists. Genes can be accessed through a gene-specific window providing reports, a dynamic map of the genomic environment, precomputed data of comparative proteomic analysis and tools for sequence analysis and downloads as described subsequently.

An important component of CandidaDB is the possibility for users to select those genomes that they wish to query from the list of all available genomes. Users can define a favourite genome, a query list of genomes and a comparative list of genomes. Through these selections, CandidaDB can be made a database focused on a favourite organism and provide comparative data for genomes of the comparative list only. The query list is used in search and comparative tools as described subsequently. Several comparative and query lists can be specified and remain accessible to registered users upon return to the database.

# ANALYSIS AND VISUALIZATION TOOLS

The migration of CandidaDB to the GenoList multigenome environment combined with the integration of nine genomes expands the possibilities for genome and proteome analysis and allows access to comparative genomics. Search options are identical to those available in the previous version of CandidaDB: the left panel of the main window allows the search by gene names and synonyms, accession numbers, text and location in the set of genomes defined by the user (favourite organism, query or comparative lists) or in all genomes present in CandidaDB. BLAST search (26) and pattern search tools are also accessible from the left panel as well as two new tools for comparative genomic analysis, FindTarget and DiffTool.

FindTarget (27) allows the user to identify genes from a given genome ('Query genome', the user-defined favourite organism) that, based on tuneable criteria (percentage of identity, E-value, etc.), are specifically present in a set of genomes ('Reference genomes', by default the user-defined query list) and, optionally, absent in another set of genomes ('Exclusion genomes', by default the user-defined comparative list). The algorithm makes use of precomputed BLASTP best hits obtained upon systematic

comparisons of all protein *versus* all proteins available in CandidaDB.

DiffTool (28) allows the identification of protein families whose components are shared by a set of organisms ('Reference genomes') as compared to another set of organisms ('Exclusion genomes'). Protein families have been pre-computed in CandidaDB using data of systematic BLASTP comparisons of every protein *versus* all proteins. Several family sets are available according to the criteria used in the clustering procedure (e.g. proteins that share at least 40, 50 or 60% sequence similarity over 80% of the protein length). Results are provided in the main window as a list of annotated protein families, each linked to the list of included proteins and a ClustalW multiple alignment (29).

Results of the different searches are displayed in the main window as gene lists, each gene being linked to a specific page that provides description, annotation and a graphical view of the genomic environment of the gene (Figure 1). Pre-computed results from comparative analysis for protein families (DiffTool) and best hits (FindTarget) and a regularly updated BLASTP comparison to the non-redundant protein databank (30) are systematically available (Figure 1). ClustalW pairwise or multiple alignments with best hits found in the genomes of the comparative list are provided. A list of bi-directional best hits (BDBH) is also provided. Additional protein features are displayed graphically showing signal peptide and membrane-spanning domains predicted using the Phobius software (31) and PFAM domains (32) (Figure 1). Direct links to relevant databases are listed in the cross-references panel (Figure 1). Tuneable, not predefined, search tools (BLAST, DiffTool, FindTarget) and sequence retrieval tools are accessible in the Analysis and Sequence tabs of this gene window, respectively.

# **CONCLUSION AND PERSPECTIVES**

The integration in a single database of a large number of genome sequences from related yeast species provides an unprecedented tool for comparative genomics of yeasts. The new version of CandidaDB aims to provide information complementary to that available at the Candida Genome Database by implementing comparative genomic tools and by providing data on functionally-relevant protein domains which were not directly available yet. Access to these data is facilitated by the use of precomputed multi-genome analysis that are normally CPU-intensive. Yet CandidaDB provides the ability to perform similar queries with user-defined parameters avoiding the limitations of these static results. The userdefined lists of genomes allow the user to limit searches and results to selected organisms, an option that will be increasingly useful when a larger number of genomes becomes available through the database.

CandidaDB is a convenient entry point for the community working on other *Candida* species than *C. albicans* since any *Candida* genome can be used as the favourite genome. It should be helpful for those who are working with genomes that are still undergoing

Control Contro Control Control Control Control Control Control Control Control Co	Collegoepilde transporter.       Searce       Mende Carl         Include added a	Accession Number orf19.2602 SDN Protein ID SLE Label -/-				
Image: market in the state in the	Description         Description         Advance         Sequence         Produce Cond           In Crodual Association         Latabal         Comparation Genomes         Produce Transmission         Produce Transm					
	Instantion         Instantion         Production         Produci	Without				
	Protecting Department in Support Ray in Status and St					
	Contractive dipper base dispersion in the set of the					
	Year	ant defects				
Type         From         To         Length (e)         Description         Procession         Procession <t< td=""><td>ppe         Pom         Pom</td></t<> <td></td>	ppe         Pom					
	Product         Diam					
	Contract the Contract of Contract o	ene				
	Constraint         Constra	See Re				
the second of the second	Contraction         Bases					
	Constraining of the second of the s					
	Constrained         Constrained <thconstrained< th=""> <thconstrained< th=""></thconstrained<></thconstrained<>					
Contraction         Contraction <thcontraction< th=""> <thcontraction< th=""></thcontraction<></thcontraction<>	Contract. Two definitions         Definitions <thdefinitions< th="">         Definitions         <thdefinition< td=""><td></td></thdefinition<></thdefinitions<>					
Contraction         Contraction <thcontraction< th=""> <thcontraction< th=""></thcontraction<></thcontraction<>	Contract. Two definitions         Definitions <thdefinitions< th="">         Definitions         <thdefinition< td=""><td></td></thdefinition<></thdefinitions<>					
Concerned         Reading Deficiency works 00 5 % inflating 00 %         Reading Mode Section Mode Mode Mode Mode Mode Mode Mode Mode	Set IF JAMLY         Randars Different Cutstra Different Cutstra Diff					
Concernence         Description         Reading Addition of the product of the produc	Default F AMBLY         Parameters         P					
Ham Data Particip Particip Particip Repairs Rep	<table-container>          Image: Product of the state of th</table-container>					
Interest     Date     Partial Statistics (S)     Date       CARSE SCA     Comparing (S)     Mathematics (S)     Mathematics (S)       Carse Sca     Scale Scal	Index Control Index Particular Reference   CALCARDE Calcardiar <t< td=""><td></td></t<>					
Addependent of the series of t	SAMARDES CASH           Same and set all all all and mannelses - see Xnahysis' lab b run methods with           Find Target         Same and the List - similarity > 30% (idebult parameters - see Xnahysis' lab b run methods with           Same and the List - similarity > 30% (idebult parameters - see Xnahysis' lab b run methods with           Same and the List - similarity > 30% (idebult parameters - see Xnahysis' lab b run methods with           Same and the List - similarity > 30% (idebult parameters - see Xnahysis' lab b run methods with           Same and the List - similarity > 30% (idebult parameters - see Xnahysis' lab b run methods with					
Bare a province of the second	Bark Histopic Score visite bis					
Sint prime in Marke Adjourned with Mit (# Bart Mark           Prima         Exercic-443         DEMACTINESTy         D. Associal CESTY7         2331         144.75         Unknown         Without Mark	Control         Multiple Algomentation with Hite (# Best Hite)           Find Target         Multiple Algomentation with Hite (# Best Hite)           Point         Multiple Algomentation with Hite (# Best Hite)         Multiple Algomentation with Hite (# Best Hite)           Point         Multiple Algomentation with Hite (# Best Hite)         Multiple Algomentation with Hite (# Best Hite)         Multiple Algomentation with Hite)         Multiple Algomentation with Hite (# Best Hite)           Point         Multiple Algomentation with Hite (# Best Hite)         Multiple Algomentation with Hite (# Best Hite)         Multiple Algomentation with Hite)         Multiple Algomentation with Hite (# Best Hite)           Point         Multiple Algomentation with Hite (# Best Hite)         Multiple Algomentation with Hite (# Best Hite)         Multiple Algomentation with Hite)           Point         Ledwagesperie         Multiple Algomentation with Hite (# Best Hite)         Multiple Algomentation with Hite)         Multiple Algomentation with Hite (# B					
ERCTCC-042         DEMACT 1992/g         D. hanseni CBS767         2311         148.47.5         United Prison           P         scale         8x121         Nuted Prison         Scale Prison	Description         Description         D. hansenif CBS767         2.33         1.484.75         Unknown           web2         8+121         333497P/8         363497P/73         3349497P/8         3349497P/8         349497P/8         349497					
No.8         P1121         NamePride         NamePrid         NamePrid         NamePride	・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・					
TULI2         OPTY         S. cervisiae 528SC         240         34.2         Productionalise dispussed on the point and maintainer methodes, also transport         Computer           P         v=0.6         v=0.4	No.         No.212C         OPT         S. cerevisiae S288C         200         86.29         Product-outplet oilgoperiside transporter of the plasma membrane; labor transporter         Product-outplet oilgoperiside transporter of the plasma membrane; labor transporter           Image: state sta	=100				
φ =0.0         φ =0.0         ψ =0.0	$4 = 0.6$ $4 = 0.6$ $3 \leq 10 \leq 10^{-1}$ $3 \leq 10 \leq 10^{-1}$ $3 \leq 10 \le 10^{-1}$ <td></td>					
And B         Senter V         Select P1         Sel	$ \begin{array}{ c c c } & \hline \begin{tabular}{ c c } & \hline \be$					
Code         Code <th< td=""><td>Conce_03873          Cguillermondi ATCC6260         2355         112.8.38         Unknown  </td><td></td></th<>	Conce_03873          Cguillermondi ATCC6260         2355         112.8.38         Unknown					
c+c0         0         Name         Na		)=89				
set@         isit@rpe@         isi	Image: Second	=100				
set@         isit@rpe@         isi	Image: Second	ptide				
etcl:         9423         9429         9429799         942999         942099         942099         942099         942099         9420999         9420999         9420999         942099999999999999999999999999999999999	$ \begin{matrix} \  \  \  \  \  \  \  \  \  \  \  \  \$					
CTING_0086          C. tropicalie MYA-3404         2352 1992.00 Uhkhoon           P         Statemplit         %Lipemplit         %Lipemplit<	CTR0_0086          C. tropicates MYA-3404         2362 1992.00 Uhkhom           CTR0_0086         S-1373         M360Pv60         SL0pemp43         M30pemp43					
velo         9-137         100 BEP / 100 Per	vector	1-30				
Image: Base 1178         1118PP/1         1119PP/1          1119PP/1	Cold         5x1178         SLIDERP/10         SLIDERP/11         SLIDERP/13	=100				
Model of the strug 20, 2007)           BDRECTONAL BEST HTS           Secsion Number         Cene Name         Organism         Similarity Percent         Similarity Percent         Overlap percent           CGUIG_03873         CGUIG_01373           CTER_00886         C. Projenism KMV-1         98.00         99.00         30.000         Cluspan="4">Cluspan="4"Cluspan="4">Cluspan="4"Cluspan="4">Cluspan="4"Cluspan="4"Cluspan="4" <th clus<="" cluspan="4" colspan="4" td=""><td>A constant webset of trun date: Aug 20, 2007)           BURCETONAL BEST HTS           Set of trun date: Aug 20, 2007)           BURCETONAL BEST HTS           Testabiol &gt; 30% similarity           Colog_03873         Coron_03873         <t< td=""><td></td></t<></td></th>	<td>A constant webset of trun date: Aug 20, 2007)           BURCETONAL BEST HTS           Set of trun date: Aug 20, 2007)           BURCETONAL BEST HTS           Testabiol &gt; 30% similarity           Colog_03873         Coron_03873         <t< td=""><td></td></t<></td>				A constant webset of trun date: Aug 20, 2007)           BURCETONAL BEST HTS           Set of trun date: Aug 20, 2007)           BURCETONAL BEST HTS           Testabiol > 30% similarity           Colog_03873         Coron_03873         Coron_03873 <t< td=""><td></td></t<>	
BURGETONAL BEST HTS           Vestbold > 30% similarity           Accession Number         Gene Name         Organism         Identity Percent         Similarity Percent         Similarity Percent         Overlap percent           CGUG_03873         CGUG_03873         CGUG_03873         C. guillemondit AFC6260         70.00         82.00         95.00           CTER_00886         CTR0_0886         CTR0_0886         C. Ropicalis MFA-3404         80.00         87.00         100.00           CKWG_01549         CAME_01549         C. albicans WO-1         98.00         99.00         30.00           YL212C         OP71         S. cerevisiae S288C         40.00         59.00         90.00           '/-         DENADC16062g         D. hansenii CBS767         66.00         79.00         100.00           CLUG_01594         CLUG_01594         C. lustanise ATCC 42720         68.00         81.00         89.00           ROTEIN FEATURES         Molecular Mass (Da) 8397.36           EC Number + -         Molecular Mass (Da) 83.97.36           EC Number + -         1         50         100         200         200         300         350         400         450         500         650         700         750 <td>BDIRECTIONAL BEST HITS           NetBOILECTIONAL BEST HITS           treashol &gt; 30% similanty         Cene Name         Organism         Identity Percent         Similarity Percent         Similarity Percent         Organism         Organism         Colspan="4"&gt;Organism         Identity Percent         Similarity Percent         Organism         Colspan="4"&gt;Organism         Organism         Identity Percent         Similarity Percent         Organism         Organism         Organism         Organism         Colspan="4"&gt;Organism         Organism         Organism</td> <td>=100</td>	BDIRECTIONAL BEST HITS           NetBOILECTIONAL BEST HITS           treashol > 30% similanty         Cene Name         Organism         Identity Percent         Similarity Percent         Similarity Percent         Organism         Organism         Colspan="4">Organism         Identity Percent         Similarity Percent         Organism         Colspan="4">Organism         Organism         Identity Percent         Similarity Percent         Organism         Organism         Organism         Organism         Colspan="4">Organism         Organism	=100				
Preshold > 30% similarity         Gene Name         Organism         Identity Percent         Similarity Percent <td>Secsion Number         Gene Number         Organism         Identity Percent         Similarly Organism         Overlap p           Accession Number         Genu 0.91873         Couro_0.91873         Couro_0.91873         Ocouro_0.91873         Ocouro_0.91873         Similarly         Accession Number         Bool         Bool         Bool         Bool         Bool         Bool         Couro_0.91873         Couro_0.91873         C. sublicalis MVA-3404         Bool         Bool         Director         Director</td> <td></td>	Secsion Number         Gene Number         Organism         Identity Percent         Similarly Organism         Overlap p           Accession Number         Genu 0.91873         Couro_0.91873         Couro_0.91873         Ocouro_0.91873         Ocouro_0.91873         Similarly         Accession Number         Bool         Bool         Bool         Bool         Bool         Bool         Couro_0.91873         Couro_0.91873         C. sublicalis MVA-3404         Bool         Bool         Director					
CGUG_03873         CDUG_93873         C. guillemondi ATCC6260         70.00         82.00         95.00           CTRG_00886         CTRG_09896         C. fopicalis M/X-3404         80.00         87.00         100.00           PICST_73580         P1CST_73580         P. stplic G85 6054         70.00         81.00         100.00           CGWG_01549         CAND_01549         G. ablicans WO-1         98.00         99.00         30.00           VILLIZC         OPT1         Scenerative S288C         40.00         59.00         90.00           -/-         DEHAOC16062g         D. hansenil C85767         66.00         78.00         100.00           LELG_00065         LELG_0045         L elongisports NRRL Y8-4239         68.00         81.00         89.00           CGUIG_01594         CLUG_01594         C. lusitaniae ATCC 42720         68.00         81.00         89.00           ROTEIN FEATURES         Molecular Mass (Da) 8397.36           EC Number -4         Solution for fig.200           1         50         100         150         200         250         300         350         400         450         500         650         700         750	CGUG_03873         CC0VG_03873         CC0VG_03873         CC0VG_03873         CC0VG_03873         C0VG         S2.00					
CTRG_00886         P stplits CBS 8054         T0.00         81.00         100.00           PLCST_73580         P1CST_73580         P1CST_73580         P1CST_73580         99.00         30.00	CTRG_00886         CTR0_00886         CTR0_00886         C. tropicalis MYA-3404         80.00         87.00         100.0           PICST_73580         P1CST_73580         P. stpits C85.6054         70.00         81.00         100.0           CAWG_01549         C. albicans WO-1         98.00         99.00         30.00           YJL212C         OPT1         S. cerevisiae S288C         40.00         59.00         99.00           -/-         DEHAOC16062g         D. hansenii C85767         66.00         79.00         100.00					
PICST_73580	PICST_73580         P1CST_73580					
VIL212C         OP71         S. cerevisiae S288C         40.00         59.00         90.00           -/-         DEMA0C18062g         D. hansenil CBS767         66.00         79.00         100.00           LELG_00055         Latta_00953         L. elongispons NRRL Y8-4239         68.00         78.00         100.00           CLUG_01594         CLUG_01934         C. lustaniae ATCC 42720         68.00         81.00         89.00           ROTEIN FEATURES         Length (a) 783 EC Number +         Molecular Mass (Da) 89397.36 tsoelectric Point 6.52         Molecular Mass (Da) 6.50         0.01         1.00           Signal Peptide 0 domains=14         Image: Provide 1         Disc.         0.02         0.02         1.00         4.00         4.50         5.00         6.00         6.50         7.00         7.50           Signal Peptide 0         Transmithane domains=14         Image: Provide 2.50         3.00         3.50         4.00         4.50         5.50         6.00         6.50         7.50           SROSE-REFERENCES         Section 1         1.50         1.00         1.50         2.50         3.00         3.50         4.00         4.50         5.50         6.00         6.50         7.50	V1/L212C         OPT1         S. cerevisiae S288C         40.00         59.00         90.00           -/-         DEHA0C18062g         D. hansenii CBS767         66.00         79.00         100.00	00				
-/-         DEMA0C18062g         D. hansenil CBS767         66.00         79.00         100.00           LELG_00055         LitLa_00045         L.elongisporus NRRL YB-4239         68.00         78.00         100.00           CLUG_01594         CLUG_01394         C.lustaniae ATCC 42720         68.00         81.00         89.00           ROTEIN FEATURES           Molecular Mass (Da) 68397.36           Source of the search of the sea	-/- DEHA0C18062g D.hansenii CBS767 66.00 79.00 100.0					
LELG_00065         LELG_0065         Leld_00963         L elongispons NRRL YB-4239         68.00         78.00         100.00           CLUG_01594         CLUG_01394         C. /ustanise ATCC 42720         68.00         81.00         89.00           ROTEIN FEATURES           Length (a)         78.3         Molecular Mass (Da)         88397.36           EC Number +         Isolecular Mass (Da)         88397.36           PFAMID         Protocolspan="4">OPT           FPG3169.6         0         OPT           FPG3169.6         0         OPT           CALE         1         Solution of the output selection of the selection of t						
CLUG_01594         CL00_01594         C. /ustaniae ATCC 42720         68.00         81.00         89.00           ROTEIN FEATURES           Length (a) 783 EC Number +         Molecular Mass (Da) 88397.36 EC Number +           PEAMIO PPAMIO PPAMIO Profiles 0         Molecular Mass (Da) 88397.36 Isoelectric Point 6.52           PEAMIO PPAMIO PAMIO PAMIO Profiles 0         Molecular Mass (Da) 88397.36 Isoelectric Point 6.52           PEAMIO Profiles 0         Point Opt Colspan="2">Opt Colspan="2">Opt Colspan="2">Opt Colspan="2">Colspan="2"Co	LELG 00065   elongisponis NRRI YR.4230 68.00 78.00 100.0					
ROTEIN FEATURES           Molecular Mass (Da) 88397.36           EC Number -4         Boolecular Mass (Da) 88397.36           PFAM 0         OPT           PFAM 0         OPT           OPT           OPT           CALE         1         50         100         150         200         250         300         450         50						

**Figure 1.** Snapshot of a gene window for the *C. albicans OPT1* gene. The gene window displays annotation data, a dynamic map of the genomic region surrounding the *OPT1* gene, access to a protein cluster including the Opt1 protein, a list of best hits identified in genomes of the comparative list with links to pairwise and multiple ClustalW alignments, a list of bi-directional best hits in other genomes available in CandidaDB, a graphical representation of predicted signal peptide, transmembrane domains and PFAM domains, and links to relevant pages in other databases. Other tabs in the gene window allow access to dynamic analysis tools and tools for sequence retrieval.

annotation. In this regard, the comparative tools available in CandidaDB can be used to refine some of the gene models provided by sequencing centers. They can also be used to focus functional genomic studies that should eventually identify gain or loss of functions that underlie the differences in pathogenicity, virulence and morphogenesis observed between the different species of the CTG clade of *Saccharomycotina*.

Other genomes of species within the CTG clade, e.g. *C. parapsilosis* and *C. dubliniensis*, have been recently sequenced and are undergoing annotation. The same is true for species of the *Saccharomycotina* that do not belong to the CTG clade. Our aim is to incorporate these genomes into CandidaDB as they become publicly available, to update sequences and annotations in a regular manner and to provide new tools for comparative and structural analysis. In particular, the incorporation in CandidaDB of a synteny visualisation tool will greatly help in the interpretation of the comparative data outputs.

#### ACKNOWLEDGEMENTS

We are grateful to Louis Jones for help in making the database publicly available. Funding to pay the Open Access publication charges for this article was provided by Institut Pasteur.

Conflict of interest statement. None declared.

#### REFERENCES

- 1. Pfaller, MA. and Diekema, DJ. (2007) Epidemiology of invasive Candidiasis: a persistent public health problem. *Clin. Microbiol. Rev.*, **20**, 133–163.
- Jones, T., Federspiel, NA., Chibana, H., Dungan, J., Kalman, S., Magee, BB., Newport, G., Thorstenson, YR., Agabian, N. *et al.* (2004) The diploid genome sequence of Candida albicans. *PNAS*, **101**, 7329–7334.
- Braun, BR., van Het Hoog, M., d'Enfert, C., Martchenko, M., Dungan, J., Kuo, A., Inglis, DO., Uhl, MA., Hogues, H. *et al.* (2005) A human-curated annotation of the Candida albicans genome. *PLoS Genet.*, 1, 36–57.
- van het Hoog, M., Rast, T., Martchenko, M., Grindle, S., Dignard, D., Hogues, H., Cuomo, C., Berriman, M., Scherer, S. *et al.* (2007) Assembly of the Candida albicans genome into sixteen supercontigs aligned on the eight chromosomes. *Genome Biology*, 8, R52.
- d'Enfert,C., Goyard,S., Rodriguez-Arnaveilhe,S., Frangeul,L., Jones,L., Tekaia,F., Bader,O., Albrecht,A., Castillo,L. *et al.* (2005) CandidaDB: a genome database for Candida albicans pathogenomics. *Nucleic Acids Res.*, 33, D353–D357.
- Arnaud, MB., Costanzo, MC., Skrzypek, MS., Binkley, G., Lane, C., Miyasato, SR. and Sherlock, G. (2005) The Candida Genome Database (CGD), a community resource for Candida albicans gene and protein information. *Nucleic Acids Res.*, 33, D358–D363.
- Arnaud, MB., Costanzo, MC., Skrzypek, MS., Shah, P., Binkley, G., Lane, C., Miyasato, SR. and Sherlock, G. (2007) Sequence resources at the Candida Genome Database. *Nucleic Acids Res.*, 35, D452–456.
- Krcmery, V. and Barnes, AJ. (2002) Non-albicans Candida spp. causing fungaemia: pathogenicity and antifungal resistance. *J. Hospital Infection*, **50**, 243–260.
- Galagan, J.E., Henn, M.R., Ma, L.-J., Cuomo, CA. and Birren, B. (2005) Genomics of the fungal kingdom: Insights into eukaryotic biology. *Genome Res.*, 15, 1620–1631.

- Dujon,B., Sherman,D., Fischer,G., Durrens,P., Casaregola,S., Lafontaine,I., de Montigny,J., Marck,C., Neuveglise,C., et al. (2004) Genome evolution in yeasts. 430, 35–44.
- Logue, ME., Wong, S., Wolfe, KH. and Butler, G. (2005) A genome sequence survey shows that the pathogenic yeast Candida parapsilosis has a defective MTLa1 allele at its mating type locus. *Eukaryot. Cell*, 4, 1009–1017.
- Jeffries, TW., Grigoriev, IV., Grimwood, J., Laplaza, JM., Aerts, A., Salamov, A., Schmutz, J., Lindquist, E., Dehal, P., et al. (2007) Genome sequence of the lignocellulose-bioconverting and xylosefermenting yeast Pichia stipitis. 25, 319–326.
- Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. 423, 241–254.
- Fischer, G., Rocha, EPC., Brunet, F.d.r, Vergassola, M. and Dujon, B. (2006) Highly Variable Rates of Genome Rearrangements between Hemiascomycetous Yeast Lineages. *PLoS Genetics*, 2, e32.
- Romov, P., Li, F., Lipke, P., Epstein, S. and Qiu, W.-G. (2006) Comparative Genomics Reveals Long, Evolutionarily Conserved, Low-Complexity Islands in Yeast Proteins. J. Mol. Evol., 63, 415–425.
- Santos, MA. and Tuite, MF. (1995) The CUG codon is decoded in vivo as serine and not leucine in Candida albicans. *Nucleic Acids Res.*, 23, 1481–1486.
- Fitzpatrick, DA., Logue, ME., Stajich, JE. and Butler, G. (2006) A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evol. Biol.*, 6, 99.
- Diezmann,S., Cox,C.J., Schonian,G., Vilgalys,RJ. and Mitchell,TG. (2004) Phylogeny and evolution of medical species of Candida and related taxa: a multigenic analysis. J. Clin. Microbiol., 42, 5624–5635.
- Goffeau, A., Barrell, BG., Bussey, H., Davis, RW., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, JD., Jacq, C. et al. (1996) Life with 6000 Genes. Science, 274, 546–567.
- Slutsky, B., Buffo, J. and Soll, D. (1985) High-frequency switching of colony morphology in Candida albicans. *Science*, 230, 666–669.
- Joly, S., Pujol, C., Schroppel, K. and Soll, D. (1996) Development of two species-specific fingerprinting probes for broad computerassisted epidemiological studies of Candida tropicalis. *J. Clin. Microbiol.*, 34, 3063–3071.
- Pappagianis, D., Collins, MS., Hector, R. and Remington, J. (1979) Development of resistance to amphotericin B in Candida lusitaniae infecting a human. *Antimicrob Agents Chemother*, 16, 123–126.
- Thanos, M., Schonian, G., Meyer, W., Schweynoch, C., Graser, Y., Mitchell, T., Presber, W. and Tietz, H. (1996) Rapid identification of Candida species by DNA fingerprinting with PCR. J. Clin. Microbiol., 34, 615–621.
- van der Walt, JP. (1966) Lodderomyces, a new genus of the Saccharomycetaceae. Antonie van Leeuwenhoek, 32, 1–5.
- Lechat, P., Hummel, L., Rousseau, S. and Moszer, I. (2008) GenoList: an integrated environment for comparative analysis of microbial genomes. *Nucl Acids Res.*, 36, D469–D474.
- Altschul,SF., Gish,W., Miller,W., Myers,EW. and Lipman,DJ. (1990) Basic local alignment search tool. J. Mol. Biol., 215, 403–410.
- Chetouani, F., Glaser, P. and Kunst, F. (2001) FindTarget: software for subtractive genome analysis. *Microbiology*, 147, 2643–2649.
- Chetouani, F., Glaser, P. and Kunst, F. (2002) DiffTool: building, visualizing and querying protein clusters. *Bioinformatics*, 18, 1143–1144.
- 29. Thompson, JD., Higgins, DG. and Gibson, TJ. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl Acids Res.*, **22**, 4673–4680.
- Wheeler, DL., Barrett, T., Benson, DA., Bryant, SH., Canese, K., Chetvernin, V., Church, DM., DiCuccio, M., Edgar, R. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucl Acids Res.*, 35, D5–D12.
- Kall,L., Krogh,A. and Sonnhammer,EL. (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics*, 21(Suppl 1), i251–i257.
- Bateman,A., Coin,L., Durbin,R., Finn,RD., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S. *et al.* (2004) The Pfam protein families database. *Nucl Acids Res.*, 32, D138–D141.