CandidaDB: a multi-genome database for Candida species and related Saccharomycotina
Tristan T. Rossignol, Pierre Lechat, Christina Cuomo, Qiandong Zeng, Ivan Moszer, Christophe d’Enfert

To cite this version:

HAL Id: pasteur-02635093
https://pasteur.hal.science/pasteur-02635093
Submitted on 31 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License
CandidaDB: a multi-genome database for Candida species and related Saccharomycotina

Tristan Rossignol¹, Pierre Lechat², Christina Cuomo³, Qiandong Zeng³, Ivan Moszer² and Christophe d’Enfert¹,*

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

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) guilliermondii, Lodderomyces elongisporus, Debaryomyces hansenii, Pichia stipitis—and the reference Saccharomyces cerevisiae genome. CandidaDB includes sequences coding for 54,170 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 user-defined 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 Saccharomycotina 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 Saccharomycoita 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) guilliermondii] 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 54,170 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.
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 CBS767 (10), an ascosporogenous species, D. hansenii strain CBS6054 (12), a xylose fermenting yeast. The new release of CandidaDB also includes the S. cerevisiae strain S288C (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 Saccharomycoymota 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

<table>
<thead>
<tr>
<th>Species</th>
<th>Strain</th>
<th>Number of proteins</th>
<th>Number of chromosomes and/or supercontigs</th>
<th>Status and release date</th>
<th>Sequencing center/Database repository</th>
<th>Database links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candida albicans</td>
<td>SC5314</td>
<td>6098</td>
<td>8</td>
<td>Draft assembly</td>
<td>CGD</td>
<td><a href="http://www.candidagenome.org/">http://www.candidagenome.org/</a></td>
</tr>
<tr>
<td></td>
<td>WO1</td>
<td>6159</td>
<td>16</td>
<td>Draft assembly</td>
<td>Broad Institute</td>
<td><a href="http://www.broad.mit.edu/annotation/">http://www.broad.mit.edu/annotation/</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>genome/candida_albicans/</td>
</tr>
<tr>
<td>Candida guilliermondii</td>
<td>ATCC6260</td>
<td>5920</td>
<td>9</td>
<td>Draft assembly</td>
<td>Broad Institute</td>
<td><a href="http://www.broad.mit.edu/annotation/">http://www.broad.mit.edu/annotation/</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>genome/candida_guilliermondii/</td>
</tr>
<tr>
<td>Candida tropicalis</td>
<td>MYA-3404</td>
<td>6258</td>
<td>23</td>
<td>Draft assembly</td>
<td>Broad Institute</td>
<td><a href="http://www.broad.mit.edu/annotation/">http://www.broad.mit.edu/annotation/</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>genome/candida_tropicalis/</td>
</tr>
<tr>
<td>Candida lusitaniae</td>
<td>ATCC42720</td>
<td>5941</td>
<td>9</td>
<td>Draft assembly</td>
<td>Broad Institute</td>
<td><a href="http://www.broad.mit.edu/annotation/">http://www.broad.mit.edu/annotation/</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>genome/candida_lusitaniae/</td>
</tr>
<tr>
<td>Debaryomyces hansenii</td>
<td>CBS767</td>
<td>6318</td>
<td>7</td>
<td>Complete</td>
<td>Genolevures</td>
<td><a href="http://cbi.labri.fr/Genolevures/eld/DEHA">http://cbi.labri.fr/Genolevures/eld/DEHA</a></td>
</tr>
<tr>
<td>Pichia stipitis</td>
<td>CBS 6054</td>
<td>5816</td>
<td>9</td>
<td>Complete</td>
<td>JGI</td>
<td><a href="http://genome.jgi-psf.org/Picst3/">http://genome.jgi-psf.org/Picst3/</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Picst3.home.html</td>
</tr>
<tr>
<td>Lodderomyces elongisporus</td>
<td>NRLL YB-4239</td>
<td>5802</td>
<td>27</td>
<td>Draft assembly</td>
<td>Broad Institute</td>
<td><a href="http://www.broad.mit.edu/annotation/">http://www.broad.mit.edu/annotation/</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>genome/lodderomyces_elongisporus/</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>S288C</td>
<td>5858</td>
<td>16</td>
<td>Complete</td>
<td>SGD</td>
<td><a href="http://www.yeastgenome.org/">http://www.yeastgenome.org/</a></td>
</tr>
<tr>
<td>Total</td>
<td>9</td>
<td>54 170</td>
<td>124</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Characteristics of the nine genomes available in the current release of CandidaDB
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, pre-computed 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 multi-genome 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 pre-computed 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 pre-computed 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 user-defined 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
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 Saccharomyces species.

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 Saccharomyces 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 synten visualization 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