RESEARCH ARTICLE

YeastIP: a database for identification and phylogeny of *Saccharomycotina* yeasts

Stéphanie Weiss[1,2], Franck Samson[3], David Navarro[4] & Serge Casaregola[1,2]

[1]INRA, UMR1319 Micalis, CIRM-Levures, Thiverval-Grignon, France; [2]AgroParisTech, UMR Micalis, Thiverval-Grignon, France; [3]INRA, UR 1077 Mathématique Informatique et Génome (MIG), Domaine de Vilvert, Jouy-en-Josas, France; and [4]INRA, UMR-1163 Biotechnologie des Champignons Filamenteux, ESIL, Marseille, France

Correspondence: Serge Casaregola, UMR1319 Micalis, INRA, CIRM-Levures, F-78850 Thiverval-Grignon, France. Tel.: +33 130 815 294; fax: +33 130 815 457; e-mail: serge.casaregola@grignon.inra.fr

Received 6 September 2012; revised 18 October 2012; accepted 23 October 2012. Final version published online 17 December 2012.

DOI: 10.1111/1567-1364.12017

Editor: Cletus Kurtzman

Keywords
yeast taxonomy; ascomycetous yeasts; multigenic analysis; sequence concatenation; phylogenetic reconstruction.

Abstract

With the advances in sequencing techniques, identification of ascomycetous yeasts to the species level and phylogeny reconstruction increasingly require curated and updated taxonomic information. A specific database with nucleotide sequences of the most common markers used for yeast taxonomy and phylogeny and a user-friendly interface allowing identification, taxonomy and phylogeny of yeasts species was developed. By 1 September 2012, the YeastIP database contained all the described *Saccharomycotina* species for which sequences used for taxonomy and phylogeny, such as D1/D2 rDNA and ITS, are available. The database interface was developed to provide a maximum of relevant information and data mining tools, including the following features: (1) the BLAST N program for the sequences of the YeastIP database; (2) easy retrieval of selected sequences; (3) display of the available markers for each selected group of species; and (4) a tool to concatenate marker sequences, including those provided by the user. The concatenation tool allows phylogeny reconstruction through a direct link to the Phylogeny.fr platform. YeastIP is thus a unique database in that it provides taxonomic information and guides users in their taxonomic analyses. YeastIP facilitates multigenic analysis to encourage good practice in ascomycetous yeast phylogeny (URL: http://genome.jouy.inra.fr/yeastip).

Introduction

The taxonomy of ascomycetous yeasts has progressed substantially over the past decade. Although phenotypic characterization is still important for species description and applied sciences, it is progressively being abandoned and replaced by molecular characterization for species delineation (Kurtzman, 2011). Kurtzman & Robnett (Kurtzman & Robnett, 1997, 1998) provided the first exhaustive ascomycetous yeast database consisting of the D1/D2 part of the large subunit of the nuclear ribosomal DNA. Species are still delineated using this marker, and it still allows the discrimination of most species, but with time and the ever-increasing documentation of the extent of biodiversity, the discriminatory value of the D1/D2 domains is slowly diminishing. Also, the D1/D2 marker is often less suitable for phylogeny. A major initiative to barcode fungi, including yeasts, has been launched, and it has been suggested to use ribosomal internal transcribed spacers (ITS) as the universal barcode marker (Seifert, 2009; Schoch *et al.*, 2012). However, recent work raises doubts about the use of ITS as the sole marker for species delineation, in view of the extent of the genetic diversity displayed by fungi (reviewed in Vralstad, 2011).

Various markers have been developed for use in phylogenetic reconstructions, including the RNA polymerase II gene (Liu *et al.*, 1999), the mitochondrial cytochrome C oxidase gene (Belloch *et al.*, 2000), and the actin gene (Daniel *et al.*, 2001; Daniel & Meyer, 2003). However, genes do not evolve similarly and often provide incongruent phylogenies when used separately, and consequently, multigenic analyses are now the preferred choice for the establishment of robust phylogenies (Kurtzman & Robnett, 2003). The Assembling the Fungal Tree of Life Project (AFTOL), aiming to generate a phylogeny of the fungal phylum, chose six loci: small subunit rRNA genes (SSU); ITS; the large subunit rRNA gene (LSU); the small subunit mitochondrial rRNA gene (mtSSU); ribosomal polymerase II largest and second largest subunit coding genes (*RPB1* and *RPB2,* respectively); and

the translation elongation factor 1-alpha gene (*TEF1*-alpha). This led to the first fungal tree based on a six-gene phylogeny (James *et al.*, 2006). Comparative genomics allowed the selection of genes on the basis of their good performance in phylogeny reconstruction in comparison with whole genome reconstruction (Kuramae *et al.*, 2007; Aguileta *et al.*, 2008), but these genes have not been widely used in a general fashion in the *Saccharomycotina* subphylum.

All marker sequences, published or just deposited, are available in the GenBank database (http://www.ncbi.nlm.nih.gov/). There are several other sources of taxonomic information. The last edition of *The Yeasts, a taxonomic study* (Kurtzman *et al.*, 2011), constitutes the reference for species description, acceptance, and the relationships between the two. The MycoBank site (http://www.mycobank.org/, Crous *et al.*, 2004) is the reference for documentation on updated taxonomical information; it is relayed by Index Fungorum (http://www.indexfungorum.org/). In addition, the CBS database (http://www.cbs.knaw.nl/) and the StrainInfo.net database (http://www.straininfo.net/, Dawyndt *et al.*, 2005) provide links between strains and associated data like taxonomy and sequences.

However, it is generally difficult for the general user or even for the taxonomist to retrieve sequences for identification or for phylogeny. Generalist databases such as NCBI and EMBL, and specialist databases like Strain.info, do not allow the retrieval of several taxonomically informative sequences at once. There is also confusion over the annotation of deposited sequences. Nilsson and coworkers (Nilsson *et al.*, 2006) have shown that up to 20% of the fungal ITS sequences in public databases have the wrong taxonomic assignment. The main problem remains the absence of simultaneous updating of taxonomical information across the various databases. These various causes make seemingly simple operations, such as authentication to the species level and phylogeny reconstruction, tedious, time consuming, and error prone.

The YeastIP database offers easy and immediate access to the type strain or representative strain sequences of each species. In addition, associated information is available in a simple, clear interface, with expert-validated sequences. YeastIP allows easy retrieval of these sequences through the use of a search tool. YeastIP was developed with the intention of helping both nonspecialists and taxonomists to establish rapid identification and phylogenetic reconstruction.

Database structure and content

A MySQL database associated with a HTML/PHP/JavaScript interface was developed as part of the YeastIP project to provide a curated database for identification and phylogeny of yeasts. The NCBI database was searched for sequences from described species belonging to the subphylum *Saccharomycotina*, and over 4348 sequences were retrieved and incorporated into YeastIP. The criteria for including these sequences were as follows:

- The sequences of type strains of all the *Saccharomycotina* species were included.
- Well-documented sequences of nontype strains were also included, especially when no or only few sequences for a given type strain were available or when the partial or complete genome sequences of nontype strains were available, for instance in the case of *Kluyveromyces lactis* (Dujon *et al.*, 2004).
- If multiple type strains are referenced in current databases (MycoBank, CBS, etc.), only the type strain of the currently recognized species was retained, as recommended by Kurtzman and collaborators (Kurtzman *et al.*, 2011).
- The sequences were also checked on the basis of quality criteria, which included their size, their proportion of informative bases and their position in the marker of interest. To ensure their traceability, all sequences included in YeastIP were previously deposited in a public database either as a single sequence with an accession number or as a complete genome sequence. In the case of sequences only available as part of a complete genome, the markers of interest were extracted and were given an arbitrary YeastIP accession number.

For the YeastIP database to be useful, each sequence file needs also to contain taxonomic information, such as clade affiliation, species name, and synonyms, for example, which are not easily available in other databases. To avoid discrepancies between the literature and the taxonomic status of sequences in databases, the name of the species associated with sequences was updated as necessary. The accepted species name in YeastIP is the name given the status 'Currently used' in the MycoBank database. For species described after the publication of '*The Yeasts, a taxonomical study*' (Kurtzman *et al.*, 2011), the accepted name in YeastIP is that given in the publication, which describes the new species. An exhaustive list of synonyms was also added to each sequence/species file as appropriate.

During the past decade, Kurtzman and his collaborators divided the yeast tree into clades. Consequently, all the species incorporated into YeastIP were attributed a clade name according to Kurtzman (2011) and to the phylogenetic information provided for new species described after 2011 as given by the authors. Note, however, that a number of species, especially in the genus *Candida*, have not been classified into an existing clade, because the phylogenetic position of these species is not sufficiently robustly established.

The markers implemented in YeastIP were chosen on the basis of several taxonomic studies (Peterson & Kurtzman, 1991; Cai *et al.*, 1996; James *et al.*, 1996, 1997; Belloch *et al.*, 2000; Daniel *et al.*, 2001; Kurtzman & Robnett, 2003; Diezmann *et al.*, 2004). The markers widely used in multi-

Table 1. Characteristics of the markers used in YeastIP

Marker	Description	Position in the gene*	Length (bp)	Number of sequences in YeastIP
LSU	Complete sequence of the large subunit 26S ribosomal RNA gene	65–3364	3300	347
D1/D2 LSU	Partial sequence of the 26S ribosomal gene comprising the D1/D2 region	65–636	570	1087
SSU	Complete sequence of the small subunit 18S ribosomal RNA gene	1–1800	1700	659
ITS	Ribosomal RNA region containing the intergenic region 1 (between 18S and 5.8S), the 5.8S ribosomal RNA gene and the intergenic region 2 (between 5.8S and 26S)	First base of ITS1 to last base of ITS2	400–600	523
mtSSU	Mitochondrial small subunit 15S ribosomal RNA gene	383–1006	400–600	340
RPB1	Partial sequence of the RNA polymerase II largest subunit coding gene	253–873	620	134
RBP2	Partial sequence of the RNA polymerase II second largest subunit coding gene	1645–2319	680–1000	163
TEF1-alpha	Partial sequence of the translation elongation factor 1-alpha coding gene	64–1190	930	374
ACT1	Partial sequence of the exon2 of the actin coding gene	405–1383	980†	226
mtCOX II	Partial sequence of mitochondrial cytochrome C oxidase subunit 2 coding gene	121–707	590	266

**S. cerevisiae* coordinates.
†540 bp in Tsui *et al.* (2008).

genic analysis are described in Table 1 and were retrieved from the NCBI for all type strains of the *Saccharomycotina* subphylum. The information provided by NCBI was conserved, and additional information from the literature has been added by the curator to complete the sequence file: current name of the species, affiliated clade if any, and an exhaustive list of synonyms. An example of a sequence file is shown in Fig. 1. A number of sequences introduced into YeastIP are already included in major phylogenic reconstructions (Kurtzman, 2003, 2006; Kurtzman & Robnett, 2007, 2010; Kurtzman *et al.*, 2007, 2008; Tsui *et al.*, 2008).

On 1 September 2012, YeastIP contained 4348 sequences representing 61 clades, 83 genera, and 1014 species of the *Saccharomycotina* subphylum. The description of the 10 markers is shown in Table 1. The database will be updated, as new species descriptions are published.

Browsing the YeastIP database

The homepage presents general information about the database content, including the number of species and the markers currently present in the database. This information is automatically updated each time new entries are implemented. The homepage also contains a 'News' section to provide information about novelties, improvements, or maintenance of the database.

The overall organization of the database is presented in Fig. 2. One of the main features of the database is the Identification tool, which identifies a strain to the species level by comparing an input sequence to the YeastIP database using BLAST N (Altschul *et al.*, 1990). The user can choose to compare the sequence of interest either to a subset containing the type strain sequences or to all sequences in YeastIP. The BLAST N tool is connected to the other tools of the database and can suggest enhancements, for instance by indicating to the user which markers are available in closely related species and could be used to infer an optimal phylogeny.

The database also provides the opportunity to search for and retrieve the sequences of markers specific to the species in the database by two methods: selecting a database entry via multiple drop-down lists and searching by keywords, as described in the following section.

Finally, the database helps the user to reconstruct phylogenies by providing information on available marker sequences for the group of species of interest and a concatenation tool to facilitate multigenic analysis.

Implemented tools

The Sequence search tool

The Sequence search tool allows the user to browse the database to retrieve sequences, either with strings in the 'keyword' field or with the clade, genus, species, and mar-

SEQUENCE 2362	STRAIN
References: gi|4038852|gb|U75427.1|PPU75427	Current name: *Barnettozyma populi* Variety: Clade: Barnettozyma Synonym cited in NCBI: Pichia populi
Definition: Pichia populi 26S ribosomal RNA gene, partial sequence Marker: LSU D1/D2 Type: T	Strain: CBS 8094
Sequence: AAACCAACAGGGATTGCCTCAGTAACGGCGAGTGAAGCGGCAAAAGCTCAAATTTGAAATCTGGTCTTCC TGCGCGGAGGCCCGAGTTGTAATTTGAAGAAGGGTTCTTGGAGAGGCGCCTTGTCTATGTTCCTTGGAAC AGGACGTCGCAGAGGGTGAGAATCCCGTTTGGCGAGGTGTGCTGATCCGTGTAAGGGCCGTTCGACGAGT CGAGTTGTTTGGGAATGCAGCTCTAAGTGGGTGGTAAATTCCATCTAAAGCTAAATATTGGCGAGAGACC GATAGCGAACAAGTACAGTGATGGAAAGATGAAAAGAACTTTGAAAAGAGAGTGAAAAAGTACGTGAAAT TGTTGAAAGGGAAGGTTATTAGATCAGACTTGGGCGGGAGCTATACTTGTCCTCTTGTAGGGCATTCACT AGCTGTCTGTTACCCGGGCCAGCATCGATTGGGGATGGCGGGAAAAAGAGCGGGGGAACGTGGCTGGGCC CTCTGTGGCGCAGTGTTTATAGCCCCTGGTTTCGCATACCGCCTGGCTCGATCGAGGACTGCGGCTTTCA TTAGCCTAGGATGCTGGCGTAATGATCTAATATCGC	Synonym: Pichia populi Hansenula populi
Number of N in sequence: 0	
Length: 596 bp	
PMID: 9850420	
Article: AUTHORS Kurtzman,C.P. and Robnett,C.J. TITLE Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences JOURNAL Antonie Van Leeuwenhoek 73 (4), 331-371 (1998)	Comment: Type species

Fig. 1. Example of a sequence file for the D1/D2 marker of *Barnettozyma populi*. Black rectangles indicate the information added by the YeastIP curator.

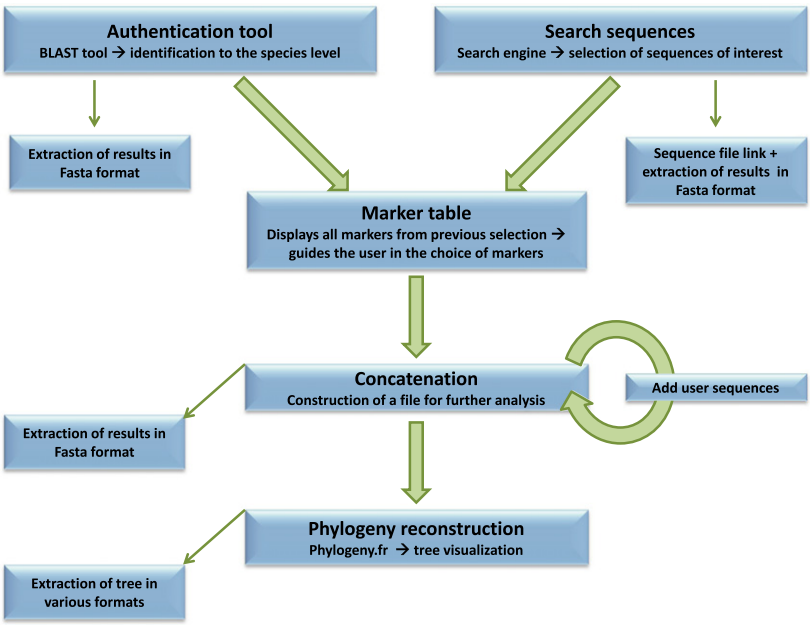

Fig. 2. Analysis pipeline in the YeastIP website.

ker names using a drop-down menu. Boolean searches can be performed on several fields.

The search by keyword allows interrogation of all fields of the database. It is particularly useful for searches starting with synonyms, as in the case of species that have undergone multiple name changes. 'Keywords' are also useful to search for sequences with either the accession number or the CBS collection number. Searches with the CBS collection number are not restricted to type strains. Because of the large number of fields scanned, the result is less specific than with clade, genus, or species names.

For example, a search for '*Saccharomyces cerevisiae*' in the keyword box gives 43 sequences; in the drop-down list, a search for the genus '*Saccharomyces*' and the species '*cerevisiae*' gives 10 sequences. These 10 sequences correspond to the 10 available markers of the *S. cerevisiae* CBS 1711^{NT} neotype strain in the database.

The Sequence search tool using clade, genus, or species will scan specific fields of the database. This type of search is recommended when the current name of the species is known by the user. The search using clade names will display all species attributed to each clade, as defined by Kurtzman (2011). Overall, it is recommended to use clade names in searches as this type of search will provide all the described related species within a clade; thus, the species belonging to the genus *Candida* and associated clades will be displayed. For example, a search with '*Lodderomyces*' as genus gives the nine sequences of the nine markers of *Lodderomyces elongisporus* present in the database. However, a search with '*Lodderomyces*' as a clade gives 90 sequences, those of *L. elongisporus* and those of the 26 *Candida* species that also belong to the *Lodderomyces* clade (Fig. 3).

Finally, the results in the form of a sequence name list are displayed with associated information such as the current name of the species, the species synonyms in NCBI, the status of the strain, the collection number, the marker name, and the length of the sequence. The accession number of each sequence provides a link to the sequence file displaying the complete information about the selected sequence (see Fig. 1 for a screen shot). Checkboxes allow the selection of sequences to be retrieved as a Fasta file or as a listing of the sequences in the first step of the concatenation tool. In the latter case, there is no need to select all the sequences of a given strain as they will all be automatically displayed in the Marker table if a sequence is selected (Fig. 4).

Concatenation/Phylogeny tools

The Identification and Sequence search tools both lead to the Concatenation/Phylogeny tools, which will process the selected sequences, as shown in Fig. 2. The Concatenation tool displays a Marker table that lists all the markers available in the database for the species selected *via* the Sequence search tool or through a blast comparison (see Fig. 4). In addition to the list of the selected species, the Marker table visualizes the presence ('YES' on a green background) or absence ('NO' on a red background) in the database of sequence markers in these strains. Thus, this table facilitates the choice of markers and strains for further phylogenetic analysis. To construct a concatenated Fasta file with the markers of each selected strain, markers and strains have simply to be selected *via* checkboxes.

The system allows for using an outgroup in the phylogenetic analysis. The sequences of some of the species most commonly used as outgroups in yeast taxonomy are provided as follows: *Pichia membranifaciens*, *Zygoascus hellenicus,* and *Schizosaccharomyces pombe* type strains (Diezmann *et al.*, 2004; Kurtzman *et al.*, 2008; Tsui *et al.*, 2008; Kurtzman & Robnett, 2010). The choice of the outgroup is an important issue in phylogenetic analysis, and these three species were found to be not entirely satisfactory as outgroups; although they carry all the markers present in the database, they can be too distantly related to some species. Therefore, it is recommended to consult the Help section for advice about choosing an appropriate outgroup to the species of interest. Note also that the users may also add their own outgroup (see below).

Once the concatenation file has been constructed, there are two possibilities: (1) the data can be retrieved as a Fasta.zip file and (2) the data can be sent to the Phylogeny.fr website (Dereeper *et al.*, 2008) for phylogenic reconstruction (see below). In addition to the ease of use of the Concatenation/Phylogeny tool, one of the main features of YeastIP is that it provides the option of including additional strains of the user's choice in the phylogenic reconstruction. All that is required is to add the relevant sequences to the concatenated file before launching the phylogenetic reconstruction. For this, the user indicates the number of strains to be added to the analysis, and this will open a new page with predetermined fields based on the markers previously selected by the user. For example, to add two strains to an *ACT1-RPB1-RPB2* concatenation, two series of three annotated boxes will be displayed. Sequences will have to be pasted and named. These sequences will be automatically concatenated and added to the previous concatenation file.

For simplicity's sake, the concatenated sequence file can then be processed with the 'One click' mode of Phylogeny.fr, which includes (1) alignment of the sequences by the MUSCLE program; (2) curation of the alignment by the GBlocks program to remove gaps and poorly conserved regions; and (3) phylogenetic reconstruction with the Maximum Likelihood program PhyML using the approximate Likelihood-Ratio Test (aLRT) for branch support, with the default settings for all these programs. The 'One Click' mode generates a phylogenetic tree that can be visualized by several programs, and/or can be downloaded in various formats (Dereeper *et al.*, 2008). To use the 'Advanced' or the 'A la carte' mode of Phylogeny.fr, the user may download the concatenated sequences as a Fasta.zip file from YeastIP. After extraction, the file may be directly processed in Phylogeny.fr or in any other sequence alignment and phylogeny programs. Finally, tree reconstruction by Phylogeny.fr uses

Select	Accession	Type	Current Name	Strain	Marker	Length	Synonym Name cited in NCBI
☑	EF120594	T	*Candida alai*	CBS 9899	LSU D1/D2	583	*Candida alai*
☑	U45776	NT	*Candida albicans*	CBS 562	LSU D1/D2	571	*Candida albicans*
☑	AY520317	T	*Candida bohiensis*	CBS 9897	LSU D1/D2	577	*Candida bohiensis*
☑	AY242341	T	*Candida buenavistaensis*	CBS 9895	LSU D1/D2	583	*Candida buenavistaensis*
☑	DQ655678	IT	*Candida chauliodes*	CBS 10157	LSU D1/D2	575	*Candida chauliodes*
☑	DQ655679	IT	*Candida corydali*	CBS 10158	LSU D1/D2	592	*Candida corydali*
☑	U57685	T	*Candida dubliniensis*	CBS 7987	LSU D1/D2	571	*Candida dubliniensis*
☑	EF120596	T	*Candida frijolesensis*	CBS 10377	LSU D1/D2	592	*Candida frijolesensis*
☑	AY520316	T	*Candida gigantensis*	CBS 9896	LSU D1/D2	581	*Candida gigantensis*
☑	AM159100	T	*Candida hyderabadensis*	CBS 10444	LSU D1/D2	584	*Candida hyderabadensis*
☑	DQ655687	T	*Candida labiduridarum*	CBS 10452	LSU D1/D2	569	*Candida labiduridarum*
☑	U45745	T	*Candida maltosa*	CBS 5611	LSU D1/D2	572	*Candida maltosa*
☑	AY497667	T	*Candida metapsilosis*	CBS 10907	LSU D1/D2	662	*Candida parapsilosis*
☑	DQ400364	T	*Candida morakotiae*	BCC 7718	LSU D1/D2	569	*Candida sp. ST-18*
☑	AF245404	T	*Candida neerlandica*	CBS 434	LSU D1/D2	570	*Candida neerlandica*
☑	FJ746056	T	*Candida orthopsilosis*	CBS 10906	LSU D1/D2	613	*Candida orthopsilosis*
☑	U45754	T	*Candida parapsilosis*	CBS 604	LSU D1/D2	570	*Candida parapsilosis*
☑	AB439258	T	*Candida prachuapensis*	CBS 11024	LSU D1/D2	570	*Candida prachuapensis*
☑	AB617978	T	*Candida sakaeoensis*	CBS 12318	LSU D1/D2	574	*Candida sp. LM078*
☑	JQ647914	T	*Candida sanyaensis*	CBS 12637	LSU D1/D2	571	*Candida sp. HN-26*
☑	AB534915	T	*Candida saraburiensis*	CBS 11696	LSU D1/D2	571	*Candida saraburiensis*
☑	U71070	T	*Candida sojae*	CBS 7871	LSU D1/D2	572	*Candida sojae*
☑	EF120599	T	*Candida tetrigidarum*	CBS 10457	LSU D1/D2	582	*Candida tetrigidarum*
☑	HM461739	T	*Candida theae*	CBS 12239	LSU D1/D2	531	*Candida sp. LCF-001*
☑	U45749	T	*Candida tropicalis*	CBS 94	LSU D1/D2	570	*Candida tropicalis*
☑	U45752	T	*Candida viswanathii*	CBS 4024	LSU D1/D2	570	*Candida viswanathii*
☑	U45763	T	*Lodderomyces elongisporus*	CBS 2605	LSU D1/D2	569	*Lodderomyces elongisporus*

Fig. 3. Search results for the *Lodderomyces* clade. The black rectangle highlights the unique representative of the genus *Lodderomyces*, whereas the *Lodderomyces* clade contains, in addition, 26 *Candida* species.

the default settings of some of the most reliable tools available. For challenging phylogenies, it is strongly recommended to download the concatenated sequence file and to adjust the options offered in the programs of the Phylogeny.fr pipeline.

Because the sequences in YeastIP originate from various sources and were directly introduced into YeastIP, the length of the sequence for a given marker may vary from strain to strain. This may affect the sequence alignment, especially when multiple markers are concatenated. To test whether the sequence length variability found in the database had an effect on sequence alignments, we tested the quality of the alignments generated by the MUSCLE program (Edgar, 2004) on many different combinations of concatenated sequences. In all cases, MUSCLE was able to produce high-quality alignments, regardless of the type of sequences and of their lengths. Similarly, we also checked the efficiency of the Gblocks program (Castresana, 2000) for the curation of alignment with large gaps. Concatenation files provided by YeastIP, which had been, or had not been, curated for large gaps between the markers, were used to generate trees. The topologies of the trees obtained with the concatenated files with and without preliminary removal of large gaps

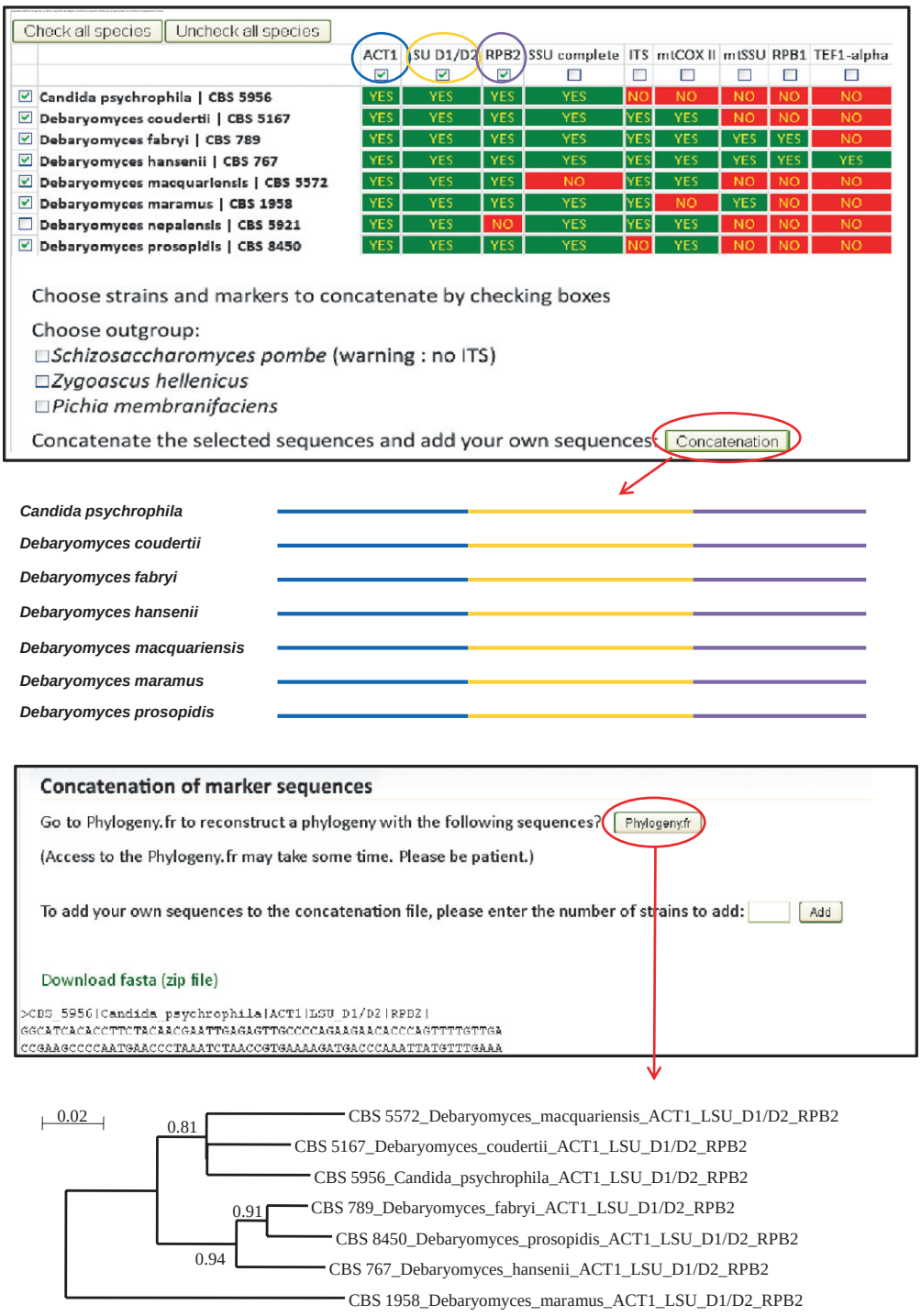

Fig. 4. Screenshot of the Marker table for some species belonging to the genus *Debaryomyces* and schematic representation of an example of concatenation.

were compared: the results were clearly similar, confirming that Gblocks could deal with large gaps in concatenated files. An example is provided in the Supporting Information, Figure S1.

Conclusion and future works

YeastIP is a database devoted to yeast identification and phylogeny by experienced taxonomists, and can

also be exploited by nonspecialists. This database contains selected sequences – LSU D1/D2, ITS, complete SSU and LSU rDNA, *ACTI*, *RPB1*, *RPB2*, *TEF1*-alpha, mtSSU, mtCOXII – and allows type strains for each species to be found easily. It contains useful tools for deducing the phylogenetic position of a strain/species on the basis of the concatenation of the sequences of several genes. YeastIP has been designed to guide the user through all the steps starting from the choice of well-adapted markers and species to the authentication or reconstruction of phylogeny. YeastIP also provides an overview of the current state of the art in taxonomy of *Saccharomycotina* yeasts and could help resolve the phylogeny of poorly studied clades by facilitating the choice of markers.

Within the framework of the European EMbaRC project (http://www.embarc.eu), YeastIP is currently dedicated on species of the *Saccharomycotina*, which is the largest yeast subphylum of the *Ascomycota*. Indeed, it includes a large number of yeasts of biotechnological and medical interest. *Saccharomyces pombe*, which belongs to the *Taphrinomycotina* subphylum, has also been included, because it is frequently used as outgroup in tree reconstruction. In the near future, other ascomycetes such as the *Taphrinaceae*, the *Protomycetaceae*, and the *Pneumocystidaceae* will also be included. We also plan to add basidiomycetous yeasts.

Acknowledgements

We gratefully acknowledge the help of the members of the CIRM-Levures. We thank the three anonymous reviewers for their help in improving the manuscript. We thank Teun Boekhout for his careful reading of the manuscript. This work has received funding from the European Community's Seventh Framework Programme (FP7, 2007-2013), Research Infrastructures action, under the grant agreement No. FP7-228310 (EMbaRC project). SW is a postdoctoral fellow in the EMbaRC project. The authors declare no conflict of interest.

References

Aguileta G, Marthey S, Chiapello H *et al.* (2008) Assessing the performance of single-copy genes for recovering robust phylogenies. *Syst Biol* **57**: 613–627.

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Belloch C, Querol A, Garcia MD & Barrio E (2000) Phylogeny of the genus *Kluyveromyces* inferred from the mitochondrial cytochrome-c oxidase II gene. *Int J Syst Evol Microbiol* **50**: 405–416.

Cai J, Roberts IN & Collins MD (1996) Phylogenetic relationships among members of the ascomycetous yeast genera *Brettanomyces*, *Debaryomyces*, *Dekkera*, and *Kluyveromyces* deduced by small-subunit rRNA gene sequences. *Int J Syst Bacteriol* **46**: 542–549.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540–552.

Crous PW, Gams W, Stalpers JA, Robert V & Stegehuis G (2004) MycoBank: an online initiative to launch mycology into the 21st century. *Stud Mycol* **50**: 19–22.

Daniel HM & Meyer W (2003) Evaluation of ribosomal RNA and actin gene sequences for the identification of ascomycetous yeasts. *Int J Food Microbiol* **86**: 61–78.

Daniel HM, Sorrell TC & Meyer W (2001) Partial sequence analysis of the actin gene and its potential for studying the phylogeny of *Candida* species and their teleomorphs. *Int J Syst Evol Microbiol* **51**: 1593–1606.

Dawyndt P, Vancanneyt M, De Meyer H & Swings J (2005) Knowledge accumulation and resolution of data inconsistencies during the integration of microbial information sources. *IEEE Trans Knowl Data Eng* **17**: 1111–1126.

Dereeper A, Guignon V, Blanc G *et al.* (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* **36**: W465–W469.

Diezmann S, Cox CJ, Schonian G, Vilgalys RJ & Mitchell TG (2004) Phylogeny and evolution of medical species of *Candida* and related taxa: a multigenic analysis. *J Clin Microbiol* **42**: 5624–5635.

Dujon B, Sherman D, Fischer G *et al.* (2004) Genome evolution in yeasts. *Nature* **430**: 35–44.

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.

James SA, Collins MD & Roberts IN (1996) Use of an rRNA internal transcribed spacer region to distinguish phylogenetically closely related species of the genera *Zygosaccharomyces* and *Torulaspora*. *Int J Syst Bacteriol* **46**: 189–194.

James SA, Cai J, Roberts IN & Collins MD (1997) A phylogenetic analysis of the genus Saccharomyces based on 18S rRNA gene sequences: description of *Saccharomyces kunashirensis* sp. nov. and *Saccharomyces martiniae* sp. nov. *Int J Syst Bacteriol* **47**: 453–460.

James TY, Kauff F, Schoch CL *et al.* (2006) Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* **443**: 818–822.

Kuramae EE, Robert V, Echavarri-Erasun C & Boekhout T (2007) Cophenetic correlation analysis as a strategy to select phylogenetically informative proteins: an example from the fungal kingdom. *BMC Evol Biol* **7**: 134.

Kurtzman CP (2003) Phylogenetic circumscription of *Saccharomyces*, *Kluyveromyces* and other members of the *Saccharomycetaceae*, and the proposal of the new genera *Lachancea*, *Nakaseomyces*, *Naumovia*, *Vanderwaltozyma* and *Zygotorulaspora*. *FEMS Yeast Res* **4**: 233–245.

Kurtzman CP (2006) New species and new combinations in the yeast genera *Kregervanrija* gen. nov., *Saturnispora* and *Candida*. *FEMS Yeast Res* **6**: 288–297.

Kurtzman CP (2011) Phylogeny of the ascomycetous yeasts and the renaming of *Pichia anomala* to *Wickerhamomyces anomalus*. *Antonie Van Leeuwenhoek* **99**: 13–23.

Kurtzman CP & Robnett CJ (1997) Identification of clinically important ascomycetous yeasts based on nucleotide divergence in the 5' end of the large-subunit (26S) ribosomal DNA gene. *J Clin Microbiol* **35**: 1216–1223.

Kurtzman CP & Robnett CJ (1998) Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. *Antonie Van Leeuwenhoek* **73**: 331–371.

Kurtzman CP & Robnett CJ (2003) Phylogenetic relationships among yeasts of the '*Saccharomyces* complex' determined from multigene sequence analyses. *FEMS Yeast Res* **3**: 417–432.

Kurtzman CP & Robnett CJ (2007) Multigene phylogenetic analysis of the *Trichomonascus*, *Wickerhamiella* and *Zygoascus* yeast clades, and the proposal of *Sugiyamaella* gen. nov. and 14 new species combinations. *FEMS Yeast Res* **7**: 141–151.

Kurtzman CP & Robnett CJ (2010) Systematics of methanol assimilating yeasts and neighboring taxa from multigene sequence analysis and the proposal of *Peterozyma* gen. nov., a new member of the Saccharomycetales. *FEMS Yeast Res* **10**: 353–361.

Kurtzman CP, Albertyn J & Basehoar-Powers E (2007) Multigene phylogenetic analysis of the Lipomycetaceae and the proposed transfer of *Zygozyma* species to *Lipomyces* and *Babjevia anomala* to *Dipodascopsis*. *FEMS Yeast Res* **7**: 1027–1034.

Kurtzman CP, Robnett CJ & Basehoar-Powers E (2008) Phylogenetic relationships among species of *Pichia*, *Issatchenkia* and *Williopsis* determined from multigene sequence analysis, and the proposal of *Barnettozyma* gen. nov., *Lindnera* gen. nov. and *Wickerhamomyces* gen. nov. *FEMS Yeast Res* **8**: 939–954.

Kurtzman CP, Fell JW & Boekhout T (2011) *The yeasts, a taxonomic study*, 5th edn. Elsevier, Amsterdam.

Liu YJ, Whelen S & Hall BD (1999) Phylogenetic relationships among ascomycetes: evidence from an RNA polymerse II subunit. *Mol Biol Evol* **16**: 1799–1808.

Nilsson RH, Ryberg M, Kristiansson E, Abarenkov K, Larsson KH & Koljalg U (2006) Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS ONE* **1**: e59.

Peterson SW & Kurtzman CP (1991) Ribosomal RNA sequence divergence among sibling species of yeasts. *Syst Appl Microbiol* **14**: 124–129.

Schoch CL, Seifert KA, Huhndorf S *et al.* (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. *P Natl Acad Sci USA* **109**: 6241–6246.

Seifert KA (2009) Progress towards DNA barcoding of fungi. *Mol Ecol Resour* **9**(suppl s1): 83–89.

Tsui CK, Daniel HM, Robert V & Meyer W (2008) Re-examining the phylogeny of clinically relevant *Candida* species and allied genera based on multigene analyses. *FEMS Yeast Res* **8**: 651–659.

Vralstad T (2011) ITS, OTUs and beyond—fungal hyper-diversity calls for supplementary solutions. *Mol Ecol* **20**: 2873–2875.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. PhyML phylogenetic tree of the species in the *Kazachstania* clade with the concatenated LSU D1/D2 and *TEF1*-alpha markers.