Journal of Biotechnology 161 (2012) 383–386

Contents lists available at SciVerse ScienceDirect

Journal of Biotechnology

journal homepage: www.elsevier.com/locate/jbiotec

Short communication

FunGene-DB: A web-based tool for Polyporales strains authentication

David Navarro[b], Anne Favel[a,*], Olivier Chabrol[c], Pierre Pontarotti[c], Mireille Haon[b], Laurence Lesage-Meessen[b]

[a] Aix Marseille Université, UMR 1163 Biotechnologie des Champignons Filamenteux, 163 avenue de Luminy, CP 925, 13288 Marseille Cedex 09, France
[b] INRA, UMR 1163 Biotechnologie des Champignons Filamenteux ESIL, 163 avenue de Luminy, CP 925, 13288 Marseille Cedex 09, France
[c] Aix Marseille Université, LATP UMR-CNRS 7353, Evolution Biologique et Modélisation, case 19, 3 place Victor Hugo, 13331 Marseille, France

ARTICLE INFO

Article history:
Received 15 March 2012
Received in revised form 20 June 2012
Accepted 25 June 2012
Available online 2 July 2012

Keywords:
Curated sequence database
Polyporales
Fungi
Molecular authentication
FunGene-DB
ITS

ABSTRACT

Polyporales are extensively studied wood-decaying fungi with applications in white and green biotechnologies and in medicinal chemistry. We developed an open-access, user-friendly, bioinformatics tool named FunGene-DB (http://www.fungene-db.org). The goal was to facilitate the molecular authentication of Polyporales strains and fruit-bodies, otherwise subjected to morphological studies. This tool includes a curated database that contains ITS1-5.8S-ITS2 rDNA genes screened through a semi-automated pipeline from the International Nucleotide Sequence Database (INSD), and the similarity search BLASTn program. Today, the web-accessible database compiles 2379 accepted sequences, among which 386 were selected as reference sequences (most often fully identified ITS sequences for which a voucher, strain or specimen, has been deposited in a public-access collection). The restriction of the database to one reference sequence per species (or per clade for species complex) allowed most often unequivocal analysis. We conclude that FunGene-DB is a promising tool for molecular authentication of Polyporales. It should be especially useful for scientists who are not expert mycologists but who need to check the identity of strains (e.g. for culture collections, for applied microbiology).

1. Introduction

Polyporales is the most representative order of saprophytic homobasidiomycetes causing wood decay. Their enzymatic potential, especially that of white-rot fungi, has been intensively investigated for applications in white and green biotechnologies (e.g. paper bleaching and pulping, biofuel production, textile dye decolourisation, bioremediation) (Mayer and Staples, 2002; Messner et al., 2003). In addition, Polyporales are also extensively studied as a source of biologically active compounds (Paterson, 2006). The sustained interest for this group of fungi makes accurate species identification essential, both for fruit-body and for isolates used in biotechnological applications. In addition, a growing number of Polyporales strains are preserved in genetic resource collections that, according to the concept of Biological Resource Centres (BRCs), must meet the highest quality assurance standards (OECD, 2001). Within this context, the identity of the deposited strains must be checked. However, this step is especially difficult for Polyporales, as several species in culture fail to develop key distinguishing growth or microscopic characteristics (Stalpers, 1978). Thus, to guarantee the identity of strains, a molecular approach would be useful as a complement to traditional methods of identification. In fact, one of the main difficulties is the reliability of the public-access DNA databases used as references, since sequences can be deposited free of control (Bridge et al., 2003; Nilsson et al., 2006). Problems are most often linked to misidentification of the original material or to poor isolation techniques. In addition, it is sometimes difficult to compare/interpret new sequences because of the increasing number of environmental sequences, the relatively poor taxonomic coverage within some groups, taxonomic uncertainty, and nomenclatural revisions (basionym, synonyms/current name).

The aim of our project was to create a curated, open-access and user-friendly rDNA database dedicated to Polyporales fungi in order to assist strain identity checking. This database, named FunGene-DB (http://www.fungene-db.org), was developed within the framework of a French BRC entitled "Centre International de Ressources Microbiennes" that hosts a collection of filamentous fungi of biotechnological interest (CIRM-CF; http://www.international.inra.fr/crb-cirm/). FunGene-DB contains high quality ITS1-5.8S-ITS2 rDNA genes selected from the

* Corresponding author. Tel.: +33 4 91828612; fax: +33 4 91828601.
E-mail addresses: david.navarro@univ-amu.fr (D. Navarro), anne.favel@univ-amu.fr (A. Favel), olivier.chabrol@univ-amu.fr (O. Chabrol), pierre.pontarotti@univ-amu.fr (P. Pontarotti), mireille.haon@univ-amu.fr (M. Haon), laurence.lesage-meessen@univ-amu.fr (L. Lesage-Meessen).

0168-1656/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jbiotec.2012.06.023

Table 1
Evolution of the FunGene-DB database content.

Number of sequences screened	Database creation: before 2009	2009	2010	2011	Total
Rejected sequences	440	98	115	185	838
Accepted sequences	882	210	373	528	1993
Reference sequences	188	51	69	78	386
Number of candidate sequences	1510	359	557	791	3217

International Nucleotide Sequence Database (INSD: GenBank, EMBL, and DDBJ). FunGene-DB includes similarity search tool through the integrated BLASTn program.

2. FunGene-DB design

FunGene-DB is a relational database built on MySQL software (version 5.0.51a) that communicates with the web interface through PHP script language (PHP version 5.2.6). A WWW server (Apache Software Foundation) giving access to the routinely updated FunGene-DB is found at http://www.fungene-db.org.

Initially, candidate ITS1-5.8S-ITS2 sequences (and the linked information) were automatically retrieved from NCBI nucleotide database using the keywords "homobasidiomycetes internal transcribed spacer". Next, because the composition of the Polyporales order used by GenBank does not agree with that of the 10th edition of the Dictionary of the Fungi (Kirk et al., 2008), sequences to be included in FunGene-DB were manually selected. Each sequence was then submitted to a multi-criteria screening procedure. Two criteria of quality were used: (i) the percentage of ambiguous sites (sequences with more than 3% of the positions reported as "n" were rejected) and (ii) the size of the ITS1-5.8S-ITS2 rDNA gene (sequences with a size less than 500 bp were rejected, except for species of high interest for the database diversity). In addition, nomenclature and taxonomy were checked. Sequences from partially identified organisms and those with invalid name according to Species Fungorum (http://www.speciesfungorum.org) were rejected. Finally, each sequence was submitted to similarity search using BLAST against all candidate sequences. The blast result (*e*-value, score, % similarity) allowed to check (i) the good enough coverage of the ITS1-5.8S-ITS2 rDNA gene and (ii) the likelihood of sequence annotation at species level.

After the screening procedure, the accepted and rejected sequences were stored in two separate databases and reasons for rejection were recorded. Rejected sequences were kept, especially in the event of a taxonomic revision. If necessary, accepted sequences were re-annotated with the current name and family name. Geographical information was systematically recorded according to a hierarchical line (continent, country and state).

For each Polyporales species, a single sequence considered as reference was selected in the accepted panel. The sub-database thus created included sequences obtained (i) directly from fruitbody for which herbarium vouchers are available or (ii) from strains deposited in an international culture collection (e.g. ATCC, CBS, DSM, FPRL, MB and MUCL), an herbarium voucher being available or not. When such sequences were not available, one environmental sequence was selected. For the species complex *Ganoderma applanatum-australe*, clade-specific sequences were selected according to Moncalvo and Buchanan (2008). For the species complex *Ganoderma lucidum*, clade-specific sequences were selected from a phylogenetic analysis of all the accepted sequences (data not shown).

3. FunGene-DB update

FunGene-DB is updated monthly (Fig. 1). New candidate sequences automatically downloaded from GenBank are processed using an additional screening step, achieved through similarity search using BLAST against the FunGene database, which allows a quick check of each new sequence to be included in FunGene-DB.

As shown in Table 1, when the database was created, 1510 candidate sequences of Polyporales fungi retrieved from INSD were screened. Among them, only 1070 sequences were accepted, and 188 sequences were selected as reference. Since 2009, the number of screened sequences has increased each year. Today, 3217 candidate sequences have been screened. They are representative of 128 genera and 500 species. Only 2379 sequences have been accepted (26% rejection), which represent 119 genera and 426 species. Among them, the sub-database compiles 386 reference sequences, representative of 116 genera and 378 species. The set of reference sequences is regularly reevaluated according to the literature.

4. Use of FunGene-DB and limits

As shown in Fig. 1, the accepted sequences dataset and the reference sub-database are web-accessible. Two possibilities are offered to users of FunGene-DB (Fig. 2). A search sequence tool allows users to browse the database either to retrieve a particular sequence (using gi or gb number or strain number as keyword) or to compile dataset sequences for further studies, e.g. phylogeny (using common keywords in taxonomic information or geographic data). Sequences selected can be recovered in FASTA format. The FunGene-DB server contains the BLASTn program v2.2.18 (Altschul et al., 1997) for similarity searches. Prior to analysis of the query sequence, users have to check in a box list whether the database includes the assumed species or not, and to make a selection ("unknown species" and "other species" are also proposed). This step was included to inform users of FunGene-DB's limitations (because of the still partial taxonomic coverage and of taxonomic uncertainties within complex groups). Among the closest matches to the query sequence, users can choose the number of sequences to be downloaded in a FASTA file. Lastly, users are prompted to make comments, especially about the choice of reference sequences, through a discussion thread.

The performance of FunGene-DB was evaluated by comparing results from BLAST (species exhibiting the highest % similarity) to those from NCBI BLAST. The test was performed with a dataset of 36 ITS sequences obtained from various, well identified strains preserved at the CIRM-CF (data not shown). Agreement between molecular and morphological identification was observed for 30 strains by using FunGene-DB versus 23 strains by using NCBI BLAST. The specificity of FunGeneDB towards Polyporales, in combination with the cleaning/re-annotation/selection step of the data allowed to avoid most of the misidentifications observed when using the raw NCBI database (data not shown). However, neither FunGeneDB- nor NCBI-BLAST identified some strains (missing taxon, taxonomic uncertainties) whose identification was thus based on morphological data only.

5. Discussion

FunGene-DB was developed to facilitate the molecular authentication of Polyporales strains and was conceived as a