Fungal Genetics and Biology 48 (2011) 353–358

Contents lists available at ScienceDirect

Fungal Genetics and Biology

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

Technological Advancement

FungiFun: A web-based application for functional categorization of fungal genes and proteins

Steffen Priebe [a,*], Jörg Linde [a,1], Daniela Albrecht [a,1], Reinhard Guthke [a,1], Axel A. Brakhage [b,c,2]

[a] Research Group Systems Biology/Bioinformatics, Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knöll-Institute (HKI), Jena, Germany
[b] Department of Molecular and Applied Microbiology, Leibniz Institute for Natural Product Research and Infection Biology – Hans-Knöll-Institute (HKI), Jena, Germany
[c] Institute of Microbiology, Friedrich Schiller University, Jena, Germany

ARTICLE INFO

Article history:
Received 22 September 2010
Accepted 1 November 2010
Available online 10 November 2010

Keywords:
Functional annotation
Fungi
Gene Ontology
FunCat
KEGG
Classification
Enrichment analysis
Candida
Aspergillus
Yeast

ABSTRACT

FungiFun assigns functional annotations to fungal genes or proteins and performs gene set enrichment analysis. Based on three different classification methods (FunCat, GO and KEGG), FungiFun categorizes genes and proteins for several fungal species on different levels of annotation detail. It is web-based and accessible to users without any programming skills. FungiFun is the first tool offering gene set enrichment analysis including the FunCat categorization. Two biological datasets for *Aspergillus fumigatus* and *Candida albicans* were analyzed using FungiFun, providing an overview of the usage and functions of the tool. FungiFun is freely accessible at https://www.omnifung.hki-jena.de/FungiFun/.

1. Introduction

High-throughput methods are frequently applied in fungal research. Using microarrays and protein arrays a multitude of biological data can be measured simultaneously. In addition, the rapid development of novel high-throughput methods like next-generation sequencing facilitates even more detailed measurements (Shendure, 2008). Global investigation of transcriptomic, proteomic or metabolomic data is essential to understand comprehensively gene regulatory mechanisms. A result of many global studies is a long list of gene or protein names. The interpretation of such gene lists is often difficult, but it is of great importance to elucidate their biological meaning. This is prerequisite to understand how a biological system reacts under a given environmental stimulus. One way of deciphering such gene or protein lists is gene set enrichment analysis (Hedegaard et al., 2009).

Several categorization approaches, which annotate specific functions of genes and their proteins, have been developed. At the same time, plenty of tools and software packages for functional categorization are on-hand. Generally, these enrichment tools can be separated into three classes: singular enrichment analysis (SEA), gene set enrichment analysis (GSEA) and modular enrichment analysis (MEA) (Huang et al., 2009a,b). FungiFun belongs to the first class, i.e. SEA, that presents the most traditional strategy for enrichment analysis. Despite the plethora of tools for well investigated species, there exists no web-based solution, which is usable without programming skills for many fungal species. The widely applied annotation tool DAVID (Huang et al., 2009a,b) does not support the common fungal gene IDs (e.g., CaO19.6385 for *Candida albicans* or AFUA_3G14490 for *Aspergillus fumigatus*). The same problem occurs with tools like GO Slimmer (Carbon et al., 2009) or FuncAssociate 2.0 (Berriz et al., 2009). GeneTrail (Backes et al., 2007) only supports *A. fumigatus* but no other fungi. In GeneCoDis2 (Carmona-Saez et al., 2007) only three yeast species are included. Other tools seem to support fungal gene IDs, but do not provide significant categories based on *p*-values (e.g., CGD GO Slim Mapper (Skrzypek et al., 2010)).

FungiFun attempts to fill this gap. The tool assigns functional annotations to a list of IDs representing fungal genes or proteins. Based on different classification methods like FunCat (Functional

---

[*] Corresponding author. Fax: +49 (0) 3641 532 2161.

E-mail addresses: steffen.priebe@hki-jena.de (S. Priebe), joerg.linde@hki-jena.de (J. Linde), daniela.albrecht@hki-jena.de (D. Albrecht), reinhard.guthke@hki-jena.de (R. Guthke), axel.brakhage@hki-jena.de (A.A. Brakhage).

[1] Fax: +49 (0) 3641 532 0803.

[2] Fax: +49 (0) 3641 532 0802.

1087-1845/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.fgb.2010.11.001

Catalog; Ruepp et al., 2004), GO (Gene Ontology; Ashburner et al., 2000) and KEGG (Kyoto Encyclopedia of Genes and Genomes; Kanehisa and Goto, 2000), FungiFun categorizes genes and proteins for currently 29 fungal species, including different human-pathogenic fungi, like *C. albicans*, *A. fumigatus* or *Candida glabrata*. These species are of major importance in the field of infection biology (Nucci and Marr, 2005; Dagenais and Keller, 2009; Brakhage et al., 2010) since they can cause superficial or allergic diseases but also invasive infections with mortality rates of up to 90% (Brakhage and Langfelder, 2002; Tekaia and Latge, 2005; Askew, 2008). Especially immunocompromised patients are at high risk to gain fungal infections, whereas candidiasis is the most frequent systemic fungal infection (Wilson et al., 2002; Brakhage and Zipfel, 2011). Beside human-pathogenic fungi, several plant-pathogenic fungi like *Botryotinia fuckeliana*, *Phaeosphaeria nodorum* or *Ustilago maydis* are also included in FungiFun. Model organisms (*Aspergillus nidulans*, *Neurospora crassa*) and industrial important species (*Aspergillus niger*, *Aspergillus terreus*, *Kluyveromyces lactis*, *Saccharomyces cerevisiae*) conclude the list of species supported by FungiFun.

Several hierarchical levels for categorization are eligible by the user to get insight into gene functions at different levels of detail. For each encountered category, an enrichment analysis is performed using Fisher's exact test. A pie chart provides an overview of the categorization. All results are available for download.

2. Material and methods

FungiFun uses a list of gene or protein names as input and performs functional classification. For each classification category, a gene set enrichment analysis is performed. The user can choose one of three different classification methods.

2.1. FunCat

FunCat is a hierarchically structured, organism-independent, flexible and scalable controlled classification system enabling the functional description of proteins from any organism. Over the last decade, FunCat has been established as a robust and stable annotation scheme that offers both, meaningful and manageable functional classification as well as ease of perception. FunCat consists of 28 main functional categories that cover general fields like cellular transport, metabolism and cellular communication/signal transduction. It exhibits a hierarchical structure with up to six levels of increasing specificity and, in total, version 2.1 includes 1362 functional categories.

Until now, there exists no enrichment analysis for the FunCat categorization. Thus, this is a novel feature within FungiFun.

2.2. GO

The goal of the GO Consortium is to produce a dynamic, controlled vocabulary that can be applied to all eukaryotes even as knowledge about the role of genes and proteins in cells is accumulating and changing. It is a collaborative effort to address the need for consistent descriptions of gene products in different databases. The GO Consortium has developed three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. GO is structured in directed acyclic graphs, where nodes represent GO terms, and edges the relationships between the terms. A GO term can have several more specialized child terms and several more general parent terms at the same time. This circumstance complicates a generalization of GO categories at a specific hierarchical level. Therefore, FungiFun does not support this option. FungiFun uses the common top terms for generalization and subdivides all specialized categories into the namespaces biological process, cellular component and molecular function.

2.3. KEGG

The Kyoto Encyclopedia of Genes and Genomes is a knowledge base for systematic analysis of gene functions, linking genomic information with higher order functional information. It offers species dependent pathway maps representing current knowledge on the molecular interaction and reaction networks. The pathway maps are subdivided into Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes, Human Diseases and Drug Development. Each of these top categories has several 2nd level categories containing specific molecular functions which are again subdivided on a third level. All in all, the KEGG pathway contains 413 functional categories.

For all methods, the categorization performed by FungiFun is based on flat files, which are downloaded from the respective servers (www.uniprot.org, http://pedant.gsf.de, www.geneontology.org, www.genome.jp/kegg) and adjusted to be easily accessible by Perl scripts. These files are updated at regular intervals, because annotations are constantly improved. The file names as well as their date of last change are stated. Consequently, the user can realize how up to date the categorization is. Due to flat file based categorization, FungiFun performs all the work in only a few seconds.

Because functional categorization of biological units is an ongoing process, not every ID submitted by the user might be annotated by a functional category. On the other hand, some IDs have no one-to-one functional mapping, since proteins accomplish multifunctional tasks. The fraction of IDs with annotations as well as the number of non-unique mappings are part of the output of FungiFun.

Results are presented by tables containing all assigned categories at the chosen and the most detailed level as well as the number of identified IDs for each category (hits). In addition, a pie chart is shown which gives an overview about the encountered categories (Fig. 3). Categories are sorted by ascending *p*-values calculated by Fisher's exact test. The *p*-value indicates the significance of the number of hits for each category in the input dataset using the number of hits for the whole genome/proteome of the organism as background. More specifically, the *p*-values describe the probability that the encountered number of hits would have been found by chance. Thus, the smaller the *p*-values the more significant the categories are. It is probable that some categories are only covered by few hits. Therefore, Fisher's exact test is used instead of a χ^2-approximation. The calculation is performed by the Text::NSP::Measures::2D::Fisher Perl package (Banerjee and Pedersen, 2003). All results are available for download as tab-separated text file.

Besides using the whole genome as background list, the user can upload a specific background list, when the analysis is based on a subset of the whole genome. In this case, not every gene for which annotation is available will be used for classification. This feature can be used for subset gene lists descending from custom microarrays. In this case, the user should use a custom-made background list, only including the subset of measured genes. The use of a specialized background list will affect the results significantly.

3. Results

3.1. FungiFun utilization

The use of FungiFun via the web-interface is self-explanatory and easy (Fig. 1). After selecting one of the currently supported species (Table 1), the user chooses a text file, which contains a list of gene or protein IDs. The user can also enter the IDs directly into a textbox on the website. For each species, some example IDs can be