Bioinformatics Advance Access published October 7, 2014

FungiFun2: A Comprehensive Online Resource for Systematic Analysis of Gene Lists from Fungal Species

Steffen Priebe [1,*], Christian Kreisel [1], Fabian Horn [1], Reinhard Guthke [1] and Jörg Linde [1*]

[1]Systems Biology / Bioinformatics; Leibniz-Institute for Natural Product Research and Infection Biology - Hans-Knöll-Institute
Beutenbergstr. 11a, 07745 Jena, Germany

Associate Editor: Dr. Janet Kelso

ABSTRACT

Summary: Systematically extracting biological meaning from omics data is a major challenge in Systems Biology. Enrichment analysis is often used to identify characteristic patterns in candidate lists. FungiFun is a user-friendly web tool for functional enrichment analysis of fungal genes and proteins. The novel tool FungiFun2 uses a completely revised data management system and thus allows enrichment analysis for 298 currently available fungal strains published in standard databases. FungiFun2 offers a modern web interface and creates interactive tables, charts and figures, which users can directly manipulate to their needs.

Availability and Implementation: FungiFun2, examples and tutorials are publicly available at https://elbe.hki-jena.de/fungifun/ .

Contact: steffen.priebe@hki-jena.de or joerg.linde@hki-jena.de

INTRODUCTION

Fungi form an extremely diverse kingdom of organisms with very different lifestyles and interesting human applications (Blackwell, 2011). Fungi are not only important to produce food but also produce bioactive compounds known as secondary metabolites (Brakhage, 2013), which are important for the pharmaceutical and chemical industries. On the other hand, there are many pathogenic fungi which destroy crops and infect humans. The growing amount of omics data from the fungal community will help to identify virulence factors as well as interesting bioactive compounds. Enrichment analysis is often applied along with omics data analysis. Here candidate genes/proteins are assigned to categories from structured vocabularies (ontologies). Afterwards, statistical tools help to identify those categories which are significantly enriched with the given candidates. These enriched categories may represent molecular functions, pathways or cellular locations most affected by the experiment. A number of easy-to-use online tools exists, e.g. YeastMine (Balakrishnan *et al.*, 2012) and are reviewed in Huang *et al.*, 2009. However, no user-friendly online tool for the systematic analysis of long candidate lists existed for most fungal species.

Our group implemented the tool FungiFun (Priebe *et al.*, 2010) supporting enrichment analysis for 28 species with a focus on fungal pathogens. In this paper, we present the novel tool FungiFun2, which allows the systematic analysis of candidate lists from all currently available fungal strains published in standard databases (Fig. 1). Users can choose from 298 strains of 240 species. For data collection, FungiFun2 uses a semi-automatic procedure which downloads gene to category associations and annotations (names, functions) from online databases. This procedure allows the database to be kept up-to-date and simplifies the addition of further species. In comparison to the previous version, which worked with flat files for annotations, FungiFun2 parses annotation into a standardised database allowing higher data connectivity and flexibility, e. g., alternative input identifiers (IDs), gene annotation, complex search queries. Finally, FungiFun2 offers a modern and user-friendly interface.

METHODS AND IMPLEMENTATION

Figure 1A illustrates the three functional ontologies which are integrated into FungiFun2, i. e., Gene Ontology (GO; Ashburner *et al.*, 2000), Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa and Goto, 2000) and Functional Catalogue (FunCat; Rüpp *et al.*, 2004). FunCat gene to category associations were downloaded from MIPS through the PEDANT database (Walter *et al.*, 2009). For GO several data source have been used: CGD (Inglis *et al.*, 2012), AspGD (Cerqueira *et al.*, 2013), SGD (Cherry *et al.*, 2012), UniProt-GOA-project at EBI and Ensembl Fungi (Kersey *et al.*, 2010). Additionally, we included GO gene to category associations by applying Blast2GO (Conesa *et al.*, 2005). To do so, proteomes were obtained either from BROAD, NCBI or in-house data. Finally, KEGG gene to pathway associations were obtained from the KEGG FTP server.

With the help of a semi-automatic procedure, all available strains in the used databases are listed. Each strain may have different data sources, where the preferred version needs to be manually selected. Afterwards, flat files are automatically downloaded and parsed into a MySQL database using Python and R scripts. These scripts guarantee that the database stays up-to-date with only small effort. Currently, ontologies formed by FunCat, KEGG and GO were downloaded from nine different sources. Primarily obtained from EBI, GO gene/protein to category association is available for 258 strains. FunCat gene/protein to category association is available for 180 strains. Finally, KEGG pathway association is available for 71 strains.

*to whom correspondence should be addressed

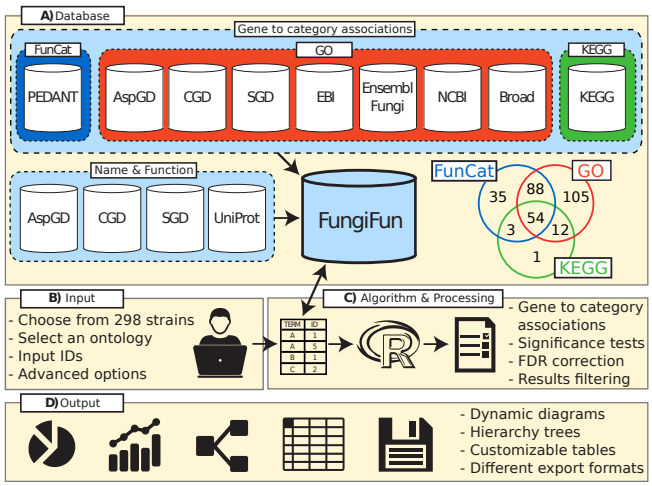

Fig. 1. Overview of FungiFun2 functionality. A) With help of a semi-automatic procedure, gene to category associations for three ontologies as well as gene names and functions are downloaded. The numbers in the Venn diagram indicate the number of available strains. B) The user selects a strain and ontology. C) On the web server, gene to category association and significance tests are performed. D) Schematic visualization of the output (dynamical figures, charts and tables).

Figure 1B illustrates main features of the user interface. To run FungiFun2, users need to chose a strain, select an ontology, supply the tool with a list of candidate IDs and choose a p-value cutoff. After strain selection, the user may check for available (alternative) IDs. Only those ontologies can be used for which annotation is currently available. Advanced options allow for alternative p-value calculations and multiple test corrections, for upload of a background list, for in/exclusion of categories, and for the selection of GO evidence codes.

Figure 1C illustrates main aspects for the calculation of enriched categories as well as results graphs and tables. On the server side, a PHP script parses user input, controls calculations of statistics, graphs and tables and finally creates data for the result page. P-values indicating the significance of the enrichment are calculated with Fisher's exact test or hypergeometric test. Multiple test correction may be performed, e.g. via FDR (Benjamini and Hochberg, 1995). The R-package `RamiGO` (Schröder *et al.*, 2013) is used to visualize significantly enriched GO categories within the GO hierarchy. Bar, pie and column charts are created with help of the JavaScript library `Highcharts`, while customizable result tables are created with JavaScript library `DataTables`.

Figure 1D illustrates parts of the results of a FungiFun2 run. Each output can be customized directly in the web interface as well as downloaded in commonly used formats. The number of enriched categories as well as the number of genes within enriched and non-enriched categories give an overview of the results. Specific pie and bar charts allow users to visualize the number of genes in the significant categories compared to the number of genes in the input list. Finally, graphs highlighting enriched categories within the hierarchies of the ontologies are available. Results are displayed in tables focusing on categories or genes which can be interactively rearranged and filtered.

ACKNOWLEDGEMENT

JL was supported by the Deutsche Forschungsgemeinschaft (DFG) CRC/Transregio 124 "Pathogenic fungi and their human host: Networks of interaction", subproject INF.

REFERENCES

Ashburner, M. *et al.* (2000). Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**(1), 25–29.

Balakrishnan, R. *et al.* (2012). Yeastmine–an integrated data warehouse for saccharomyces cerevisiae data as a multipurpose tool-kit. *Database (Oxford)*, **2012**, bar062.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.

Blackwell, M. (2011). The fungi: 1, 2, 3 ... 5.1 million species? *Am J Bot*, **98**(3), 426–438.

Brakhage, A. A. (2013). Regulation of fungal secondary metabolism. *Nat Rev Microbiol*, **11**(1), 21–32.

Cerqueira, G. C. *et al.* (2013). The aspergillus genome database: multispecies curation and incorporation of rna-seq data to improve structural gene annotations. *Nucleic Acids Res*.

Cherry, J. M. *et al.* (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res*, **40**(Database issue), D700–D705.

Conesa, A. *et al.* (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**(18), 3674–3676.

Huang, D. W. *et al.* (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, **37**(1), 1–13.

Inglis, D. O. *et al.* (2012). The Candida genome database incorporates multiple Candida species: multispecies search and analysis tools with curated gene and protein information for Candida albicans and Candida glabrata. *Nucleic Acids Res*, **40**(Database issue), D667–D674.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, **28**(1), 27–30.

Kersey, P. J. *et al.* (2010). Ensembl Genomes: extending Ensembl across the taxonomic space. *Nucleic Acids Res*, **38**(Database issue), D563–D569.

Priebe, S. *et al.* (2010). FungiFun: A web-based application for functional categorization of fungal genes and proteins. *Fungal Genet Biol*, **48**, 353–358.

Rüpp, A. *et al.* (2004). The funcat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res*, **32**(18), 5539–5545.

Schröder, M. S. *et al.* (2013). RamiGO: an R/Bioconductor package providing an AmiGO visualize interface. *Bioinformatics*, **29**(5), 666–668.

Walter, M. C. *et al.* (2009). PEDANT covers all complete RefSeq genomes. *Nucleic Acids Res*, **37**(Database issue), D408–D411.