D456–D458 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj026

FGDB: a comprehensive fungal genome resource on the plant pathogen *Fusarium graminearum*

Ulrich Güldener[1,3,*], Gertrud Mannhaupt[2], Martin Münsterkötter[3], Dirk Haase[3], Matthias Oesterheld[3], Volker Stümpflen[3], Hans-Werner Mewes[1,3] and Gerhard Adam[4]

[1]Chair of Genome Oriented Bioinformatics, Center of Life and Food Science, Technische Universität München, D-85350 Freising-Weihenstephan, Germany, [2]Department of Organismic Interactions, Max-Planck-Institute for Terrestrial Microbiology, D-35043 Marburg, Germany, [3]GSF National Research Center for Environment and Health, Institute for Bioinformatics, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany and [4]Department of Applied Plant Sciences and Plant Biotechnology, Institute of Applied Genetics and Cell Biology, BOKU-University of Natural Resources and Applied Life Sciences, Muthgasse 18, A-1190 Vienna, Austria

Received August 11, 2005; Revised and Accepted September 21, 2005

ABSTRACT

The MIPS *Fusarium graminearum* Genome Database (FGDB) is a comprehensive genome database on one of the most devastating fungal plant pathogens of wheat and barley. FGDB provides information on two gene sets independently derived by automated annotation of the *F.graminearum* genome sequence. A complete manually revised gene set will be completed within the near future. The initial results of systematic manual correction of gene calls are already part of the current gene set. The database can be accessed to retrieve information from bioinformatics analyses and functional classifications of the proteins. The data are also organized in the well established MIPS catalogs and novel query techniques are available to search the data. The comprehensive set of gene calls was also used for the design of an Affymetrix GeneChip. The resource is accessible on http://mips.gsf.de/genre/proj/fusarium/.

INTRODUCTION

The ascomycete *Fusarium graminearum* (anamorph *Gibberella zeae*) causes plant diseases of world-wide economic importance. *Fusarium* head blight of wheat, barley and other small grain cereals, and ear rot of maize lead to reduced yield and, most importantly, to contamination of agricultural products with *Fusarium* mycotoxins. To protect consumers, the European Commission has recently set maximum tolerated levels for the *F.graminearum* mycotoxins deoxynivalenol and zearalenone for various food commodities [COMMISSION REGULATION (EC) No 856/2005].

Fusarium genomics aims to resolve some of these serious environmental health problems. Insights into the virulence mechanisms of the broad range pathogen and antagonistic defense mechanisms in host plants is expected to allow for knowledge-based breeding of crop plants with improved resistance. The identification of fungal virulence factors should also be useful for the development of chemical compounds targeting the most relevant mechanisms, which could shift the balance in favor of plant resistance. The MIPS *Fusarium graminearum* Genome Database (FGDB; http://mips.gsf.de/genre/proj/fusarium/) provides a comprehensive resource for the international research community. It features a continuously updated set of the estimated 14 000 genes and downstream analysis in an ongoing gene validation process.

SOURCE DATA AND CONTENT OF THE DATABASE

The source data for FGDB were provided by the *F.graminearum* sequencing project at the Broad Institute, which is supported by the National Research Initiative which is part of the US Department of Agriculture's (USDA's) Cooperative State Research Education and Extension Service. All data are based on the provided 511 contig sequences (http://www.broad.mit.edu/). Gene models were imported from (i) the Broad gene set of 11 640 automatically predicted putative genes (Calhoun annotation system), (ii) the MIPS draft gene call set (13 938 genes) identified by the program FGENESH (www.softberry.com) with a matrix trained on fungal sequences of diverse origin (*Ustilago maydis*, *Schizosaccharomyces pombe*, and others), and (iii) the MIPS pipeline of manually processed gene calls, currently about 600 altered or new calls (prefix fg12). The latter uses several gene prediction programs [Fgenesh, GENEMARK (1),

*To whom correspondence should be addressed. Tel: +49 89 3187 3579; Fax: +49 89 3187 3585; Email: u.gueldener@gsf.de

GENEFINDER (P. Green, unpublished data), GENSCAN (2) in combination with different matrices], information on matching expressed sequence tag (EST) sequences (3) and BlastX analysis. Human experts integrate the data to find the most reliable gene model (4). The automatically predicted gene calls are either replaced by manually corrected calls if they are found to be wrong, or tagged as 'manually processed' if they are analyzed but not altered.

During this manual gene modeling and correction procedure it appeared that the MIPS draft gene call set performed significantly better than the Broad set. The MIPS draft gene calls have fewer erroneously fused gene calls and very few falsely added short 5′ and 3′ exons. Thus, the MIPS gene set resulted in a higher number of predicted calls. Manual processing of the automatically extracted gene calls first focused on the main groups of genes relevant in plant–pathogen interaction. During the manual annotation process it appeared that roughly 65% of the Broad calls and only 25% of the MIPS draft calls are most probably incorrect in their exon structure.

The database may hold different gene calls for a certain locus, but only one is valid at a time. Currently the genome is represented by 14 086 valid genes, 900 of which are either new or manually processed. The corresponding proteins are classified into six classes: (i) class 'known protein' are those corresponding to previously characterized *F.graminearum* genes, whose protein sequences are available in the public databases (43 proteins), (ii) 'strong similarity to known protein' (amino acid identity $\geqslant$60%, 1797 proteins), (iii) 'similarity to known protein' (amino acid identity between 40 and 60%, 4304 proteins), (iv) 'similarity to unknown protein' (3173 proteins), (v) 'strong similarity to EST' (318 proteins) and (vi) 'no similarity' (4451 proteins). Protein titles were assigned by mapping titles from previously annotated proteins. Iterative blasts were performed extracting titles by homology (identity >40%) to manually annotated datasets (*Neurospora crassa* and *Saccharomyces cerevisiae*), a subset of the SwissProt database and a non-redundant protein database with varying stringencies. All proteins are characterized by the PEDANT system (5), which provides an analysis based on a suite of bioinformatics tools for the assignment of functional and structural attributes. The results are imported into the core database enabling convenient queries and browsing (see below). Beside several protein structural analyses, e.g. detection of low complexity regions, non-globular regions, coiled coil regions and transmembrane region prediction, the PEDANT system also provides results of functional classification (6). In addition, the results of HMM search against the Pfam database and a set of results of PSI-BLAST similarity searches against various databases, e.g. a non-redundant protein database (MIPS compiled), a database of protein sequences with known 3D structure (PDB), MIPS databases of proteins with manually assigned functional categories, the SCOP database of protein domains, the COG database and InterPro are provided. For the assignment of a protein to a functional category, the coverage ratio (identity * length/hit length) $\geqslant$ 25% was used (7). Unsuitable functional categories were manually revised, e.g. plant specific categories. All other method results were not manually validated and default cut-off scores were applied.

The comprehensive set of predicted genes has already been used for the design and development of an Affymetrix GeneChip (funded by the USDA, expected to become publicly available at the end of 2005). For all genes, corresponding GeneChip probes are listed in the entry report and visualized in the DNA view of FGDB.

In order to explore the similarity space of single proteins as well as on a proteome scale, the similarity matrix of protein sequences (SIMAP) is used (8). SIMAP is a MIPS developed resource, which currently provides a pre-calculated all-against-all comparison of 3.5 million proteins including the proteomes of 35 fungal species. The best coverage ratio levels of certain taxonomical areas like 'Fungi', 'Ascomycota', 'Bacteria' or 'Mammalia' are provided in each single FGDB entry with a link to a complete list of homologs in this area. Queries on these taxonomic homology data, like 'show me all FGDB entries with homology to bacteria >50% and homology to mammalian <30%', are possible. Additionally for each entry, the closest homolog among all taxons is extracted and organized in a best hit taxonomy catalog.

Syntenic regions are extracted by sequence similarity to several fungi and bacteria. The coverage ratio (identities * overlap/length, with respect to the longer protein; minimum coverage ratio 20%) was used together with the localization and orientation of the open reading frames (ORFs) on the contigs. The results can be explored starting on the entry or contig level.

In addition to protein coding sequences as FGDB entries, information on 306 tRNAs, identified using tRNAscanSE (9) and 165 mapped genetic markers (10) are integrated.

IMPLEMENTATION AND DATABASE STRUCTURE

The data resource is realized in Genome Research Environment (GenRE), a multi-tier architecture, following a component-based approach by Güldener *et al*. (11). GenRE complies to the J2EE specification for distributed components. This approach allows seamless integration with the other fungal genome resources at MIPS.

The data model follows a generic specification for a fungal genome database at MIPS. In contrast to other resources, the model allows storage of abundant protein-related information derived from experimental and bioinformatics analyses. It contains tables storing results from manual and predicted functional annotations, domain information from InterPro, or basic physical features about the genes and their derived proteins.

The component-oriented approach combined with the generic specification provides an infrastructure for comparative genomics in fungal organisms. One of the main advantages is the network-transparent access to distributed components. For example, similarity between proteins in different organisms can be efficiently retrieved with a component accessing the SIMAP located at the Technical University of Munich (8).

RETRIEVAL OF INFORMATION

FGDB offers two alternative ways to retrieve information. The first one requires minimal user input and provides information on proteins, genes and contigs (with direct access on the sidebar). Additionally, proteins are categorized according to

available classification schemes such as the MIPS Functional Catalog (6), Enzyme Classification, Taxonomy of the closest homologs, InterPro, Protein Classification and TMHMM as implemented for the Comprehensive Yeast Genome Database (12). This allows web-based navigation through the categories.

The second alternative is an advanced query interface similar to the Entrez service at NCBI (13). Important database fields are indexed for customizable searches such that no knowledge of the underlying data structures is required. Full-text queries on all indexed information are possible and can be combined using logical operators (http://mips.gsf.de/genre/proj/fusarium/Search/Gise/). An example query for all valid proteins, which are transport ATPases (FunCat Category 20.03.22) and actually have an ATPase domain (using InterPro Description), would simply be '1[ENI] 20.03.22[FCC]*ATPase*[IPD]' instead of a complicated native database query, which would involve several table joins.

In addition, the functional distribution of a pasted or uploaded list of genes can be retrieved using a feature of the FunCatDB (6). A BLAST service has been established (http://mips.gsf.de/genre/proj/fusarium/Search/) to search for homologous contig, orf and protein sequences, that also covers DNA sequences, which are not part of the contig assembly.

DOWNLOAD/LINKS

Complete sets of *F.graminearum* sequences and annotation can be downloaded from ftp://ftpmips.gsf.de/fusarium/. These include lists of genetic elements and the contig sequences of the automatically predicted gene sets as well as the current valid gene set. The functional classification can be found on ftp://ftpmips.gsf.de/fusarium/catalogues/.

FUTURE DIRECTIONS

The ongoing process of manual curation of the gene calls should achieve a completely, manually revised gene set within the near future. With the upcoming sequencing of further *Fusarium* species (*Fusarium oxysporum* and *Fusarium verticillioides*, funded through the USDA/NSF Microbial Sequencing Program), this process will be substantially facilitated by comparative genomics. It will be possible to distinguish genuine ORFs from annotation artifacts and to assign start codons and intron/exon boundaries in a more reliable manner. Also the elucidation of conserved non-coding elements will benefit from the integration of further genome datasets. Affymetrix GeneChip expression analysis results will be integrated in FGDB as well as in PLexDB, a community resource for plant and plant–pathogen microarrays. Interconnecting links of both resources will facilitate access to a variety of analysis methods including datasets on the host plants. As FGDB allows integration of data from the research community, e.g. disruption phenotypes or experimental evidences for exon/intron structure (cDNA sequencing), further input is welcome. Our database schema allows additional attributes, e.g. all data will be tagged with PubMed ID and evidence (experiment type, source of data like 'journal' or 'personal communication').

CONCLUSIONS

FGDB is a comprehensive resource on the fungal plant pathogen *F.graminearum*. It will be continuously enhanced by ongoing gene call processing aided by using sequence data of further *Fusarium* species. The established database structure and retrieval techniques facilitate a user friendly web portal and blue print for upcoming fungal genome databases.

ACKNOWLEDGEMENTS

We thank Louise Riley for critical reading of the manuscript and the *Fusarium* research community for input on gene models and annotation details. This work was supported by the Austrian genome programme GEN-AU (bm:bwk—Federal Ministry for Education, Science and Culture, GZ 200.051/6-VI/1/2001), the Impuls- und Vernetzungsfonds der Helmholtz-Gemeinschaft and a grant of the German Federal Ministry of Education and Research (BMBF) within the BFAM framework (031U112C/212C). Funding to pay the Open Access publication charges for this article was provided by the Austrian genome programme GEN-AU.

Conflict of interest statement. None declared.

REFERENCES

1. Besemer,J. and Borodovsky,M. (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res*., **33**, W451–W454.
2. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*., **268**, 78–94.
3. Trail,F., Xu,J.R., San Miguel,P., Halgren,R.G. and Kistler,H.C. (2003) Analysis of expressed sequence tags from *Gibberella zeae* (anamorph *Fusarium graminearum*). *Fungal Genet. Biol*., **38**, 187–197.
4. Schulte,U., Becker,I., Mewes,H.W. and Mannhaupt,G. (2002) Large scale analysis of sequences from *Neurospora crassa*. *J. Biotechnol*., **94**, 3–13.
5. Riley,M.L., Schmidt,T., Wagner,C., Mewes,H.W. and Frishman,D. (2005) The PEDANT genome database in 2005. *Nucleic Acids. Res*., **33**, D308–D310.
6. Ruepp,A., Zollner,A., Maier,D., Albermann,K., Hani,J., Mokrejs,M., Tetko,I., Güldener,U., Mannhaupt,G., Münsterkötter,M. *et al*. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res*., **32**, 5539–5545.
7. Wilson,C.A., Kreychman,J. and Gerstein,M. (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol*., **297**, 233–249.
8. Arnold,R., Rattei,T., Tischler,P., Truong,M.D., Stümpflen,V. and Mewes,W. (2005) SIMAP—the similarity matrix of proteins. *Bioinformatics*, **21** Suppl. 2, ii42–ii46.
9. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*., **25**, 955–964.
10. Gale,L.R., Bryant,J., Calvo,S., Giese,H., Katan,T., O'Donnell,K., Suga,H., Taga,M., Usgaard,T.R., Ward,T.J. *et al*. (2005) Chromosome complement of the fungal plant pathogen *Fusarium graminearum* based on genetic and physical mapping and cytological observations. *Genetics*, doi:10.1534/genetics.105.044842.
11. Güldener,U., Münsterkötter,M., Oesterheld,M., Pager,P., Ruepp,A., Mewes,H.W. and Stümpflen,V. (2006) MPact: The MIPS Protein Interaction Resource on Yeast. *Nucleic Acids Res*., **34**, D436–D441.
12. Güldener,U., Münsterkötter,M., Kastenmüller,G., Strack,N., Van Helden,J., Lemer,C., Richelles,J., Wodak,S.J., Garcia-Martinez,J., Perez-Ortin,J.E. *et al*. (2005) CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Res*., **33**, D364–D368.
13. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., Di Cuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids. Res*., **33**, D39–D45.