Nucleic Acids Research Advance Access published November 4, 2010

Nucleic Acids Research, 2010, 1–3
doi:10.1093/nar/gkq1016

FGDB: revisiting the genome annotation of the plant pathogen *Fusarium graminearum*

Philip Wong[1], Mathias Walter[1], Wanseon Lee[1], Gertrud Mannhaupt[1,2], Martin Münsterkötter[1], Hans-Werner Mewes[1,3], Gerhard Adam[4] and Ulrich Güldener[1,*]

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

Received September 14, 2010; Revised October 7, 2010; Accepted October 10, 2010

ABSTRACT

The MIPS *Fusarium graminearum* Genome Database (FGDB) was established as a comprehensive genome database on one of the most devastating fungal plant pathogens of wheat, barley and maize. The current version of FGDB v3.1 provides information on the full manually revised gene set based on the Broad Institute assembly FG3 genome sequence. The results of gene prediction tools were integrated with the help of comparative data on related species to result in a set of 13.718 annotated protein coding genes. This rigorous approach involved adding or modifying gene models and represents a coding sequence gold standard for the genus *Fusarium*. The gene loci improvements results in 2461 genes which either are new or have different structures compared to the Broad Institute assembly 3 gene set. Moreover the database serves as a convenient entry point to explore expression data results and to obtain information on the Affymetrix GeneChip probe sets. The resource is accessible on http://mips.gsf.de/genre/proj/FGDB/.

INTRODUCTION

The ascomycete *Fusarium graminearum* (anamorph *Gibberella zeae*) is the causal agent of several plant diseases of world-wide economic importance (1). *Fusarium* head blight of cereals and *Fusarium* ear rot of maize lead to severe yield losses and quality problems. Most importantly, mycotoxins (2) produced by the pathogen contaminate infected plant material and derived food and feed products leading to a health risk. To protect consumers and to avoid a negative impact on farm animals, maximum tolerated levels for *Fusarium* toxins have been enacted in many countries and costly mycotoxin monitoring programs were implemented. The most sustainable solution to the problem seems to be breeding resistant plants. Yet, this is difficult, because the molecular basis of quantitative resistance differences are not understood (3). The pathogen has a very broad host range and seems to be able to suppress plant defense responses in ways that are currently not understood or to a very limited extent (4). The elucidation of fungal virulence mechanisms and the identification of virulence genes that can be targeted by breeding or biotechnological approaches is the main goal of a large research community. As a first step in the development of genomics tools for *F. graminearum* and as a basis for functional genomics approaches, the full genome sequence of one *F. graminearum* strain was determined (5).

The setup of the first version of the FGDB (6) was supported by a project funded by the Austrian genome initiative GEN-AU and was based on the first genome assembly. It already focused on manual improvements of gene calls. The intuitive user interface allowed access to the data through various search and browsing methods. Input from the research community enhanced the annotation effort and established the resource as a key tool for *F. graminearum* genomics (5).

The current FGDB v3.1 (http://mips.gsf.de/genre/proj/FGDB/) aims to provide a comprehensive resource for the international research community based on the latest assembly of the genome sequence and on a manually revisited set of 13.718 genes, 319 tRNAs and genetic markers with a detailed functional annotation and bioinformatic analysis. In addition, the database was

*To whom correspondence should be addressed. Tel: +49 89 3187 3582; Fax: +49 89 3187 3585; Email: u.gueldener@helmholtz-muenchen.de

expanded to provide convenient access to available GeneChip expression data.

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 being part of the US Department of Agriculture's (USDA's) Cooperative State Research Education and Extension Service. The current content of FGDB v3.1 is based on the Broad assembly 3 resulting in 31 supercontigs (7). The Broad Institute used the previous FGDB version 1 with its manually revised gene calls to improve their current gene set. Based on this set, all gene loci in FGDB v3.1 were re-annotated using a pipeline including (i) Fgenesh with different matrices (www.softberry.com); (ii) GeneMark-ES (8); (iii) Augustus with ESTs, precedingly annotated *Fusarium* models and/or *Neurospora crassa* protein sequences as training data or as hints for the predicted model structure (9); and (iv) EST data as well as Blastx data of related *Fusarium* species (*F. verticillioides*, *F. oxysporum* and *F. solani*). The different models were displayed in GBrowse (10) allowing comprehensive manual validation of the coding sequences (CDSs). The best fitting model per locus was selected manually and in case for required changes, respective gene calls were manually corrected using Apollo (11). The gene identifiers have been retained unchanged from the Broad FG3 gene set if the model was identical. All altered (1770) or newly added gene calls (691) are named FGSG_15xxx and above. The outdated draft identifiers used for the Affymetrix GeneChip design (fgdxx-xxx, 13 938 genes) (12) and the corresponding FG1 identifiers (fgxxxxx, 11 640 genes) are listed as alias in the entry pages and are linked to Pedant databases for details (13).

The ORF data and resulting protein sequences are imported in the Pedant system for a detailed functional and structural bioinformatic analysis. The core results are re-imported into FGDB for convenient display and indexing. The Pedant analysis details are inter-linked with each FGDB entry. The assembly 1 data were used for the design of an Affymetrix GeneChip (12). The single probes were mapped on the supercontigs using Blat at 100% identity. Probe sets corresponding to gene loci are searchable and visualized in the GBrowse viewer. The initial expression analysis results are integrated for a brief overview on the expression of single genes. Similarity based data (e.g. homology between protein pairs) is retrieved from and interlinked to the Similarity Matrix of Proteins (SIMAP), which is updated on a monthly interval (14).

Comparison of the Broad FG3 and FGDB v3.1 annotated gene sets indicate that 11 257 genes (82%) are exactly the same in terms of exon/intron structure. A total of 2461 genes in the Broad set either have a different structure or are absent from FGDB. A total of 2056 genes in FGDB either have a different structure or are absent from the Broad data. With the evidence of protein similarity to related species, 26 genes in the Broad set have been split into two or more genes in FGDB while 147 genes in FGDB were merged from two or more genes of the Broad set. Overall, FGDB v3.1 contains 383 more introns than the Broad set, with a decrease in mean intron length from 83.4 to 76.6 nt. Both annotation sets have ~65% of genes annotated with at least one putative InterPro domain (15). The average number of domains annotated per gene for both Broad and FGDB is ~1.7. As judged by confirmation of introns by available ESTs, both Broad and FGDB are of similar quality indicating that the validation of gene calls by available EST data was similarly efficient for both pipelines.

There are 103, 55 and 1651 proteins predicted only in FGDB, only in Broad and in both annotation sets as part of the secretory pathway [TargetP, RC < 4 (16)], respectively. In particular, both Broad and FGDB models now enable secretion prediction of FGSG_17357 (related to inorganic pyrophosphatase IPP1) and FGSG_12369 (related to catalase 2) as identified previously in an extracellular proteomics study (17) on models without SignalP signals (18). In addition, FGDB predictions help confirm the secretory pathway membership of hypothetical protein FGSG_16372 as identified in that study.

RETRIEVAL OF INFORMATION

The database interface provides basic search options on the sidebar which allows full text search across gene codes, gene symbols and gene description. In addition, the annotation catalogs FunCat (19), Enzyme Class (20), InterPro (14) and Protein Class are browsable. The advanced search page offers access also to invalid gene models which disagree with known evidences, details on the GeneChip data like probe and probe set names and their location (12), tables on tRNAs and a customizable table on protein molecular weights and isoelectric points. The ORF / contig DNA and protein sequences are searchable by Blast.

The single entry page of a gene locus lists information on outdated gene models, alias names and protein classification (six classes from known to hypothetical). Beside physical features like contig coordinates, molecular weight, etc., the hierarchical, functional classification FunCat (19) and EC-number classes (20) as well as InterPro IDs (15) and TargetP (16) results are provided. SIMAP based protein homology data can be retrieved using links grouped by NCBI-based taxonomic categories.

The Pedant links shown in the individual gene records forward to the respective Pedant report pages including alternative views on the DNA level as well as a graphic protein feature view. A small contig pictogram on the right side of each individual gene report page is linked to a GBrowse view allowing graphical browsing of genes, GeneChip probes, EST data and outdated gene models on their corresponding contigs.

To get a brief overview on the initial expression analysis data (12,21–23) for single genes, the 'Expression Data' link placed below the contig pictogram provides a brief description of experiments and presents the expression data for all matching probe sets. In addition, a more comprehensive overview of the most recent expression data is

provided by a link to the 'PLEXdb GeneOscilloScope' (24). The advanced query option (Index Search) on the left panel can be used to retrieve a list of the current FGDB entries based on complex queries including InterPro domains, TargetP results and e.g. probe set names (e.g. "fgd122-100_at"[pgs]|"fgd122-620_at"[pgs]). For this purpose, the major database fields are indexed which allows a fast and combined 'index search' (http://mips.gsf.de/genre/proj/FGDB/Search/Gise/).

DOWNLOAD/LINKS

The data can be downloaded from ftp://ftpmips.gsf.de/FGDB/. Beside the protein, contig and chromosome sequence file in fasta format the ORF data is provided in gff3 format. Functional data like FunCat, TargetP and InterPro are accessible in tab-delimited files.

CONCLUSIONS AND FUTURE DIRECTIONS

The FGDB v3.1 is a comprehensive resource on the fungal plant pathogen *Fusarium graminearum* and facilitates a user friendly access to gene structure and functional data. Protein homology-based data from public genomes is routinely updated. Although the ORFeome is completely revised in this version, updates on single gene structures are likely to come as new sequence data of further *F. graminearum* strains and closely related species or EST data are available in future. We encourage any input of additional evidence to further improve the gene set and overall annotation of the genome. Submitted links to gene specific publications, contact information on existing mutation strains and other details will also be included.

FUNDING

Austrian Science Fund FWF (special research project Fusarium, F3702 and F3705). Funding for open access charge: Helmholtz Zentrum München, German Research Center for Environmental Health, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany.

Conflict of interest statement. None declared.

REFERENCES

1. Goswami,R.S. and Kistler,H.C. (2004) Heading for disaster: Fusarium graminearum on cereal crops. *Mol. Plant Pathol.*, **5**, 515–525.
2. Desjardins,A.E. (2006) *Fusarium Mycotoxins: Chemistry, Genetics and Biology American Phytopathological Society*., http://www.apsnet.org (19 October 2010, date last accessed).
3. Buerstmayr,H., Ban,T. and Anderson,J.A. (2009) QTL mapping and marker-assisted selection for Fusarium head blight resistance in wheat: a review. *Plant Breeding*, **128**, 1–26.
4. Walter,S., Nicholson,P. and Doohan,F.M. (2010) Action and reaction of host and pathogen during Fusarium head blight disease. *New Phytol.*, **185**, 54–66.
5. Cuomo,C.A., Güldener,U., Xu,J., Trail,F., Turgeon,B.G., Di Pietro,A., Walton,J.D., Ma,L., Baker,S.E., Rep,M. *et al.* (2007) The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization. *Science*, **317**, 1400–1402.
6. Güldener,U., Mannhaupt,G., Münsterkötter,M., Haase,D., Oesterheld,M., Stümpflen,V., Mewes,H. and Adam,G. (2006) FGDB: a comprehensive fungal genome resource on the plant pathogen Fusarium graminearum. *Nucleic Acids Res.*, **34**, D456–D458.
7. Ma,L., van der Does,H.C., Borkovich,K.A., Coleman,J.J., Daboussi,M., Di Pietro,A., Dufresne,M., Freitag,M., Grabherr,M., Henrissat,B. *et al.* (2010) Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. *Nature*, **464**, 367–373.
8. Ter-Hovhannisyan,V., Lomsadze,A., Chernoff,Y.O. and Borodovsky,M. (2008) Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.*, **18**, 1979–1990.
9. Stanke,M., Keller,O., Gunduz,I., Hayes,A., Waack,S. and Morgenstern,B. (2006) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.*, **34**, W435–W439.
10. Donlin,M.J. (2009) Using the Generic Genome Browser (GBrowse). *Curr. Protoc. Bioinformatics*, **28**, 9.9.1–9.9.25.
11. Ed,L., Nomi,H., Mark,G., Raymond,C. and Suzanna,L. (2009) Apollo: a community resource for genome annotation editing. *Bioinformatics*, **25**, 1836–1837.
12. Güldener,U., Seong,K., Boddu,J., Cho,S., Trail,F., Xu,J., Adam,G., Mewes,H., Muehlbauer,G.J. and Kistler,H.C. (2006) Development of a Fusarium graminearum Affymetrix GeneChip for profiling fungal gene expression in vitro and in planta. *Fungal Genet. Biol.*, **43**, 316–325.
13. Walter,M.C., Rattei,T., Arnold,R., Güldener,U., Münsterkötter,M., Nenova,K., Kastenmüller,G., Tischler,P., Wölling,A., Volz,A. *et al.* (2009) PEDANT covers all complete RefSeq genomes. *Nucleic Acids Res.*, **37**, D408–D411.
14. Rattei,T., Tischler,P., Arnold,R., Hamberger,F., Krebs,J., Krumsiek,J., Wachinger,B., Stümpflen,V. and Mewes,W. (2008) SIMAP–structuring the network of protein similarities. *Nucleic Acids Res.*, **36**, D289–D292.
15. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
16. Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
17. Paper,J.M., Scott-Craig,J.S., Adhikari,N.D., Cuomo,C.A. and Walton,J.D. (2007) Comparative proteomics of extracellular proteins in vitro and in planta from the pathogenic fungus Fusarium graminearum. *Proteomics*, **7**, 3171–3183.
18. Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.
19. 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.
20. Gasteiger,E., Gattiker,A., Hoogland,C., Ivanyi,I., Appel,R.D. and Bairoch,A. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, **31**, 3784–3788.
21. Hallen,H.E., Huebner,M., Shiu,S., Güldener,U. and Trail,F. (2007) Gene expression shifts during perithecium development in Gibberella zeae (anamorph Fusarium graminearum), with particular emphasis on ion transport proteins. *Fungal Genet. Biol.*, **44**, 1146–1156.
22. Seong,K., Zhao,X., Xu,J., Güldener,U. and Kistler,H.C. (2008) Conidial germination in the filamentous fungus Fusarium graminearum. *Fungal Genet. Biol.*, **45**, 389–399.
23. Hallen,H.E. and Trail,F. (2008) The L-type calcium ion channel cch1 affects ascospore discharge and mycelial growth in the filamentous fungus Gibberella zeae (anamorph Fusarium graminearum). *Eukaryotic Cell*, **7**, 415–424.
24. Wise,R.P., Caldo,R.A., Hong,L., Shen,L., Cannon,E. and Dickerson,J.A. (2007) BarleyBase/PLEXdb. *Methods Mol. Biol.*, **406**, 347–363.