Nucleic Acids Research, 2006, Vol. 34, Database issue ***D465–D470***
doi:10.1093/nar/gkj119

Phytophthora functional genomics database (PFGD): functional genomics of *phytophthora*–plant interactions

Kamal Gajendran, Michael D. Gonzales, Andrew Farmer, Eric Archuleta, Joe Win[1], Mark E. Waugh and Sophien Kamoun[1,*]

National Center for Genome Resources, Santa Fe, NM, USA and [1]Department of Plant Pathology, The Ohio State University-Ohio Agricultural Research and Development Center, Wooster, OH, USA

Received August 15, 2005; Revised and Accepted October 19, 2005

ABSTRACT

The Phytophthora Functional Genomics Database (PFGD; http://www.pfgd.org), developed by the National Center for Genome Resources in collaboration with The Ohio State University-Ohio Agricultural Research and Development Center (OSU-OARDC), is a publicly accessible information resource for *Phytophthora*–plant interaction research. PFGD contains transcript, genomic, gene expression and functional assay data for *Phytophthora infestans*, which causes late blight of potato, and *Phytophthora sojae*, which affects soybeans. Automated analyses are performed on all sequence data, including consensus sequences derived from clustered and assembled expressed sequence tags. The PFGD search filter interface allows intuitive navigation of transcript and genomic data organized by library and derived queries using modifiers, annotation keywords or sequence names. BLAST services are provided for libraries built from the transcript and genomic sequences. Transcript data visualization tools include Quality Screening, Multiple Sequence Alignment and Features and Annotations viewers. A genomic browser that supports comparative analysis via novel dynamic functional annotation comparisons is also provided. PFGD is integrated with the Solanaceae Genomics Database (SolGD; http://www.solgd.org) to help provide insight into the mechanisms of infection and resistance, specifically as they relate to the genus *Phytophthora* pathogens and their plant hosts.

INTRODUCTION

Oomycetes, in particular *Phytophthora* spp., comprise a unique branch of destructive eukaryotic plant pathogens that are responsible for causing a number of world's most devastating diseases of dicot plants (1). Not only are these diseases difficult to manage, but also they cause enormous economic damage to important crop species such as potato, tomato and soybean, as well as environmental damage in natural ecosystems. These plant pathogenic microbes have the remarkable ability to manipulate biochemical, physiological and morphological processes in their host plants via effector molecules (2). In collaboration with The Ohio State University-Ohio Agricultural Research and Development Center (OSU-OARDC), we have constructed the Phytophthora Functional Genomics Database (PFGD; http://www.pfgd.org). We are experimentally characterizing effector gene and protein sequences, gene expression patterns, biological activity, as well as cellular responses and localization during infection. PFGD is a publicly accessible, web-based information resource designed to capture these heterogeneous data in a useful and intuitive way for biological researchers. PFGD interrelates functional assays, transcript, genomic and expression analysis. PFGD was built upon data formerly available from the Phytophthora Genome Consortium (3) and the Syngenta Phytophthora Consortium (4), as well as from all publicly available *Phytophthora infestans* and *Phytophthora sojae* transcript data and *P.infestans* genomic data all of which are analyzed and annotated using NCGR's XGI (Genome Initiative for species X) automated computational pipeline (http://www.ncgr.org/xgi) (5). PFGD integrates with NCGR's Solanaceae Genomics Database (SolGD; http://www.solgd.org) to explore plant–pathogen interactions. SolGD hosts the expression studies of Solanaceae response to *P.infestans* isolates and will soon incorporate expression profiles for unique

*To whom correspondence should be addressed. Tel: +1 330 263 3847; Fax: +1 330 263 3841; Email: kamoun.1@osu.edu
Present address:
Mark E. Waugh, Los Alamos National Laboratory, Los Alamos, NM, USA

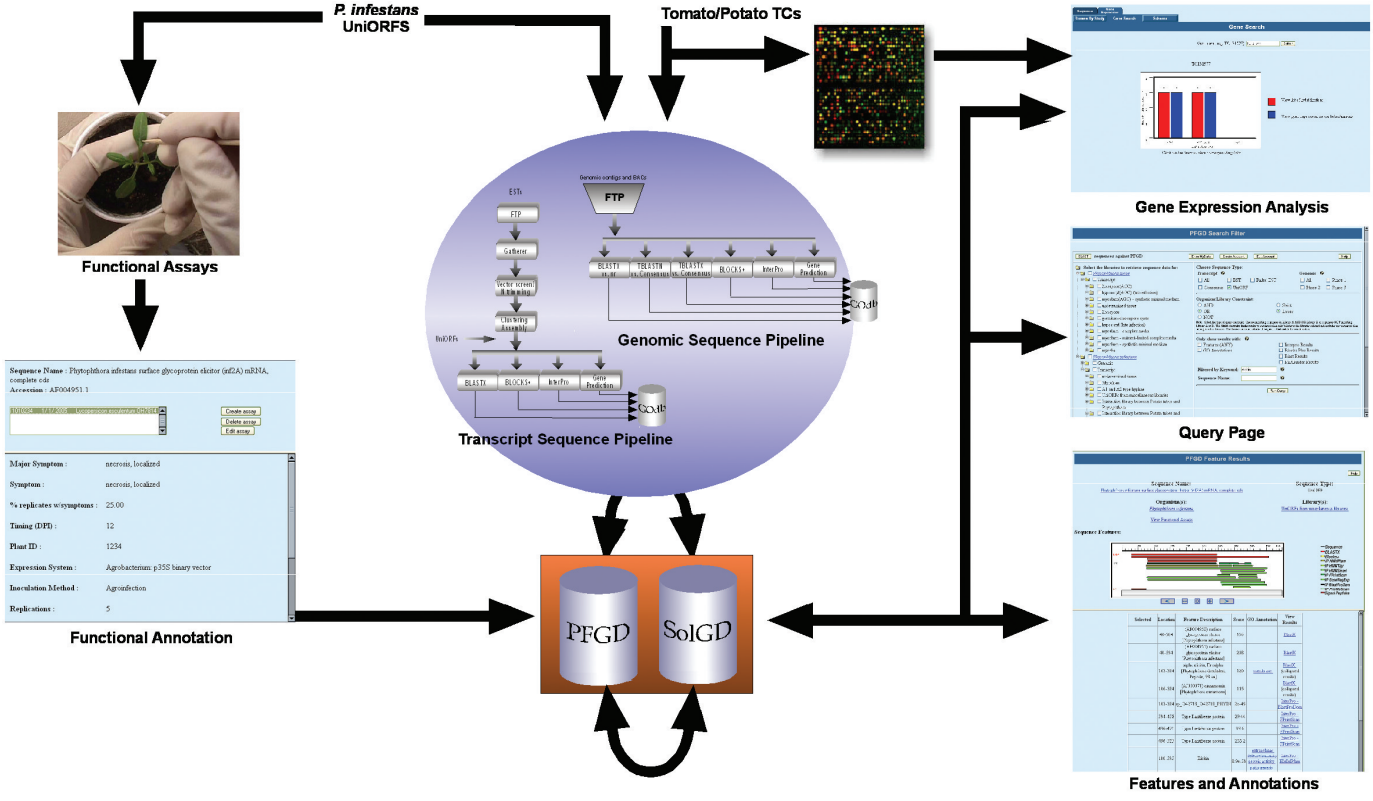

Figure 1. PFGD and SolGD architecture. Information flow starts with DNA sequences from *P.infestans* or solanaceous plants (such as tomato and potato). DNA sequences are annotated either by transcript sequence pipeline or genomic sequence pipeline developed at NCGR. Annotation data for *P.infestans* are deposited in the PFGD, and for tomato and potato, in the Solanaceae Genome Database (SolGD). Both databases have similar user interfaces accessed through the Internet via a web browser, starting with a sequence search on the Query Page. Users can then inspect annotated data by activating appropriate feature links associated with a sequence. PFGD contains a facility for users to enter functional analysis data of UniORF sequences that are used in their experiments, while SolGD is furnished with microarray data from *Phytophthora*–Solanaceae interactions. PFGD and SolGD are interlinked so that users can seamlessly browse the data between the two databases.

open reading frames (UniORFs) in comparable plant–pathogen systems as data become available. These federated databases are designed to provide significant insight into key molecular processes regulating an economically important pathosystem and a useful and accessible computational platform for the study of disease resistance in crop plants. We also point the readers to two other resources, Oomycete Molecular Genetics Research Collaboration Network (http://pmgn.vbi.vt.edu) and the DOE-JGI eukaryotic genomes (http://genome.jgi-psf.org/euk_cur1.html) that provide complementary data to that of PFGD.

PFGD ARCHITECTURE

PFGD integrates transcript, genomic, gene expression and functional assay data from a number of sources and allows users to access and compare data via a multifaceted web interface (Figure 1). PFGD has a modular architecture, which can best be described as a local federation of databases. This allows each module to be updated independently reflecting the reality of differential data generation in the scientific community. For example, the functional assays are curated in real-time as data become available, whereas the transcript module is updated based on expressed sequence tag (EST) count thresholds set for each species. A core database module has been developed to encapsulate data common to more than one module. For instance, the same user login information can be used to access MyData modules of both PFGD and SolGD.

All publicly available transcript and genomic data from *P.infestans* and *P.sojae* have been analyzed using NCGR's XGI system. Analysis is done within each species and the results can be used in cross-species comparisons to identify species-specific gene sequences. Comparative transcript data are used in genomic analyses by aligning gene sequences to genomic contigs to validate *ab initio* gene predictions. Functional assay experiments performed on host plants with full-length *P.infestans* ORFs (UniORFs) are manually curated at OSU and uploaded to PFGD. Transcript data, TIGR Tentative Consensus or consensus sequences generated by XGI, for Solanaceae host plants are analyzed using the XGI system and imported into SolGD. Gene expression experiments performed on Solanaceae with *P.infestans* UniORFs will be imported into SolGD as they become available and will be tightly integrated with PFGD. Various user interface tools to query and visualize the above-mentioned data content are provided.

DATA CONTENT

PFGD makes use of the XGI system for automated analysis and annotation of transcript data, including ESTs, consensus sequences and full-length ORFs (UniORF) as well as genomic

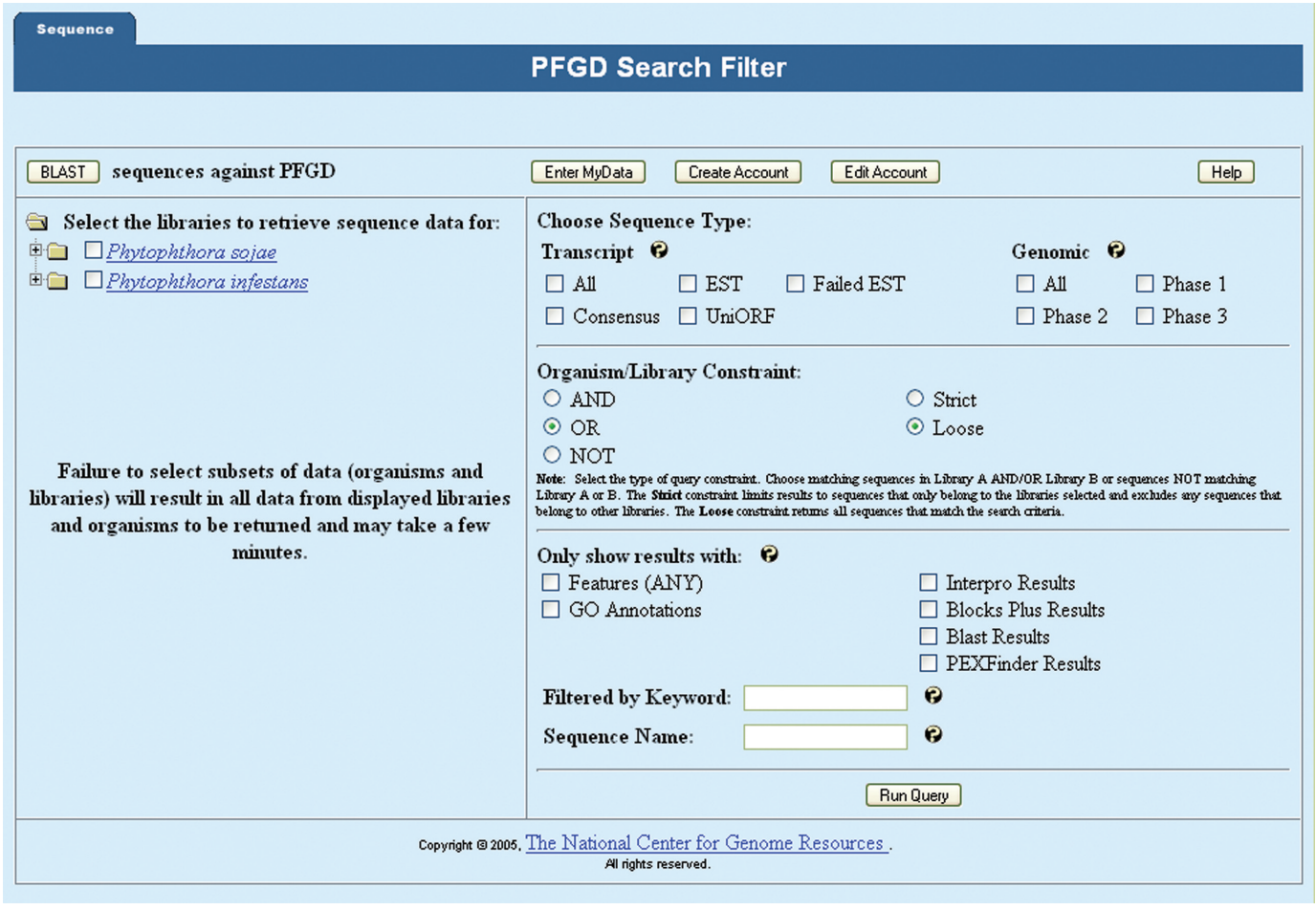

Figure 2. PFGD query page. To access the database, users start by selecting desired origins and type of sequences (e.g. organisms and transcript/genomic/treatment conditions). These sequences can then be searched using BLAST by activating the 'BLAST' button at the top left side of the page, or by selecting the features associated with the sequence and/or entering the keyword or sequence name on the right panel of the query page. The search options can be constrained by using Boolean logic, such as AND, OR and NOT. This page also contains options for users to set up their own account to store data they have searched. The sequences matching the criteria set on this page are presented to the users when they activate the 'Run Query' button.

data. Sequence data are processed through a series of analysis operations, each operation building upon the results of the previous stage. Functional assay experiments are manually curated whereas expression data are imported in an automated bulk upload.

Transcript data

The PFGD transcript data content consists of quality-screened EST and derived consensus sequences for *P.infestans* and *P.sojae* as well as UniORF sequences for *P.infestans*. Using the XGI transcript pipeline, raw public EST and cDNA data are gathered from NCBI and analyzed. Where available, quality scores for EST sequences are incorporated into the database for use by downstream analysis components. Detailed metadata concerning sequence origins such as submitting organization, organism, library details and cloning methodology, are captured and are viewable in the PFGD interface.

Before being used in analyses, raw EST data are screened for quality. Screening operations include removal of most common vector sequences, poly (A/T) trimming, N-trimming, adapter/linker removal and poor-quality read trimming. Vector screening and adapter/linker screening remove sequence contamination of the insert that typically arises as part of the cloning process. In addition, the fidelity of a sequence read typically degenerates toward the end of the sequence, resulting in errors in base calling which are trimmed out as part of this process. Finally, low-complexity sequences represented by polyadenylated regions can produce many false positive matches in downstream analyses. The end result of the quality screens is a high-quality 'approved' sequence that is then deposited in the database. An EST that has failed the XGI vector screen analysis for one or more reasons is not included in subsequent analyses, but may still be inspected through the interface as a failed EST.

Approved EST data are clustered using Phrap (http://www.phrap.org), which performs clustering and contig assembly to produce consensus sequences; these aggregate the high-quality sequence information of their member ESTs and are used in all downstream analyses. Consensus sequences are analyzed using NCBI's blastx algorithm (6) to search for potential homologues against NCBI's nr database (7). BlimpsSearcher analysis (8) against the Blocks+ database (9)

is used to identify protein blocks. InterProScan (10) is run to integrate results from a variety of protein motif analysis tools using the InterPro database (11). Each of these analyses is followed by a stage that uses the results to associate Gene Ontology (GO) terms (12) with the sequences. Pexfinder (Phytophthora Extracellular Proteins Finder) (13), co-developed by NCGR and OSU-OARDC, based on SignalP (14) has also been incorporated. Pexfinder predicts proteins secreted through the pathogen plasma membrane based on scanning for signal peptides in cDNA sequences. The results of the pipeline analyses are stored in the PFGD database and are updated periodically depending on the number of public sequences available for analysis.

UniORFs are full-length *P.infestans* cDNA sequences derived from select consensus of interest based on sequence analysis and annotation from PFGD. UniORFs are used in functional assays on potato and tomato as well as in gene expression experiments using microarrays based on tomato or potato consensus sequences annotated and stored in SolGD. UniORFs are also annotated using the above mentioned post-clustering analyses.

Genomic data

PFGD genomic data are processed using the XGI genomic pipeline. Public sequence information is gathered from NCBI's High Throughput Genomic (HTG) division (7) for the species of interest. HTG sequences are from large-scale genome sequencing centers and are submitted as in-process assemblies in various stages of completeness, often containing two or more contigs; PFGD does not assemble its genomic data, but takes data from GenBank as submitted by the sequencing centers. Each genomic sequence is then separated into its constituent contigs and analyzed in pieces using a sliding window of length 10 000 with an overlap of 3000 bp. These pieces are processed using blastx (6) against NCBI's nr database (7), and with blastn and tblastx (6) against the consensus sequences produced by the PFGD transcript pipeline. BlimpsSearcher (8) and InterProScan (10) are used as described previously. Analysis results that are in common between overlapping pieces are merged before being stored. GenScan (15), which performs *ab initio* gene prediction on the genomic sequences, is run on the complete contig sequences, providing the opportunity to define and compare the genes and exon–intron organization of the sequences. The results of the genomic pipeline are stored in the PFGD database and are updated periodically depending on the number of sequences available for analysis.

Functional assay data

Functional assay experiments performed on host plants using full-length *P.infestans* ORFs (UniORFs) are manually curated by the OSU-OARDC into PFGD via the functional assay curation interface. A user account management system allows only privileged users to create or edit experiments, although all PFGD users can view the experiments already curated in PFGD.

Gene expression data

An extensible MIAME compliant data model, designed in such a way that it integrates well with the existing data model for transcript and genomic data, stores the gene expression data. Tomato gene expression experiments have been imported into this gene expression component using an automated upload mechanism. Sequence and gene expression data are tightly integrated through the user interface.

USER INTERFACE

The PFGD web interface provides a powerful set of search tools to access, compare, and save transcript, genomic, gene expression and functional assay data. Sequence and analysis data are logically organized and searchable in a variety of ways. The PFGD interface allows intuitive navigation of all publicly available transcript and genomic data (Figure 2). A list of libraries is provided in a hierarchical structure based on the organism, sequence type, library name and organization which produced the library for user selection. Precise delineation of library selection can be accomplished by using the boolean logic operators (AND, OR, NOT) in conjunction with scope delimiters (STRICT, LOOSE) thus enabling virtual Northerns and *in silico* subtractions. More sophisticated queries using keyword searches on features and GO annotations are also supported. The interface also gives users the ability to search for results and annotation based on specific types of analysis tool output (e.g. only retrieve sequences that have InterPro results) or combinations of output (retrieve only sequences with both Blocks+ and InterPro results). Access to functional assays is provided for each curated UniORF.

The interface presents data in a variety of formats, including graphical depictions of sequences decorated with their predicted features, multiple sequence alignments (MSAs) with highlighted sequence variants and detailed reports of analysis operations. These annotations can also be downloaded in batches using the interface.

Sequence details

The sequence detail views capture all information relevant at the individual sequence level, and other information can be accessed through this display. This includes quality trimming, base composition, as well as sequence metadata and clustering information. For Consensus and UniORF sequences, it is also a portal to all analyses run on the sequence.

Features and annotations

The Features and Annotation (F&A) view displays all data available for Consensus sequences. Links to sequence details, MSAs, EST membership data as well as library and organism metadata are provided. The F&A also gives a graphical presentation of the analysis results linked directly to GO (12) annotations where appropriate. In addition, the direct text output from the analysis results can be viewed by following the appropriate links.

Multiple sequence alignments

Clustering results are summarized as MSAs. MSAs are associated with any transcript sequence in the database (both Consensus sequences, and ESTs) with mismatches between EST and consensus sequences highlighted in red.

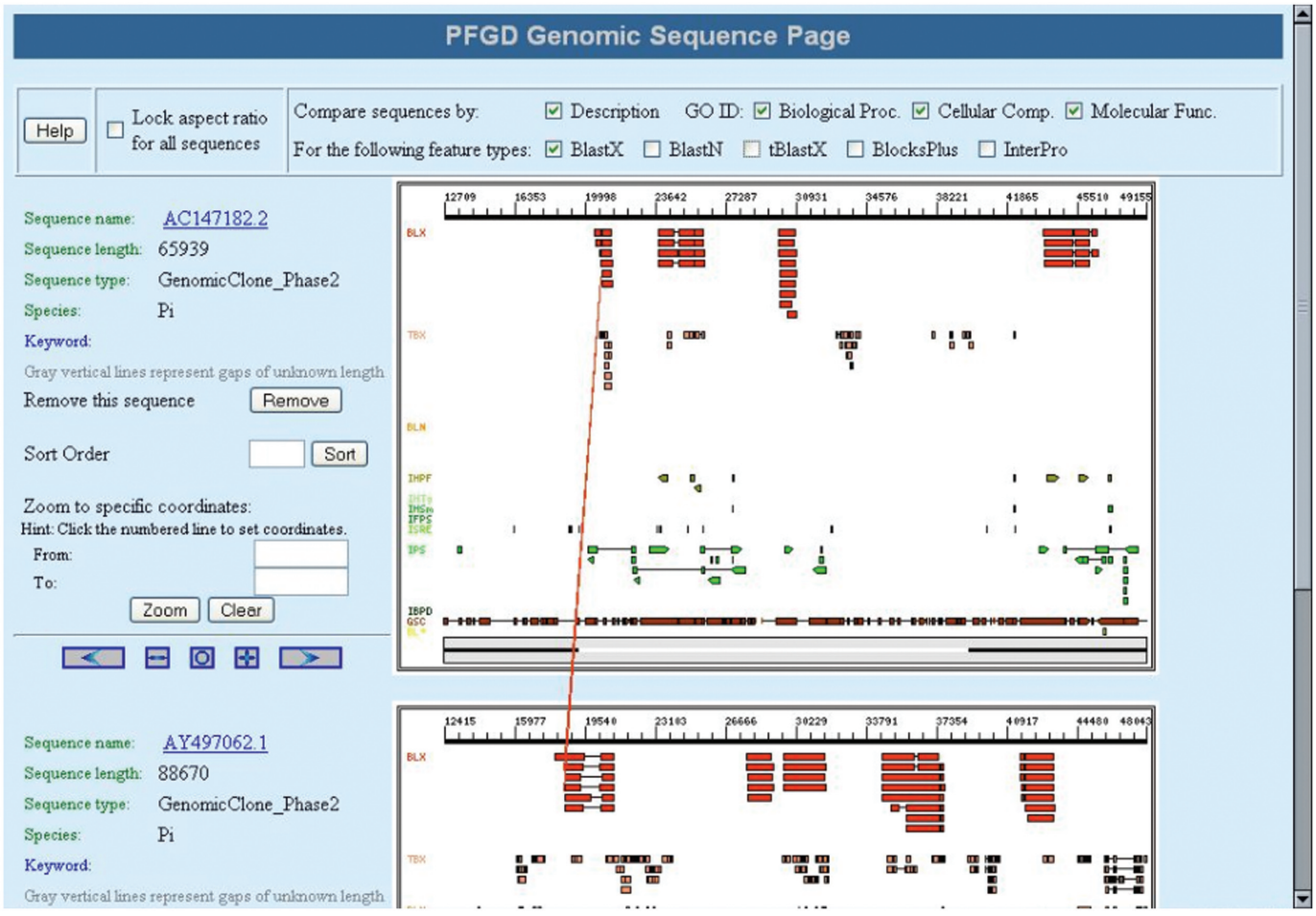

Figure 3. The Comparative Functional Genomics Browser. This screenshot shows a comparison of two bacterial artificial chromosome clones containing genomic sequences of *P.infestans*. In this case, the comparison was performed by selecting all annotations (check boxes in the top panel) obtained via Blastx analysis of the sequences for the two BAC clones. Related regions are shown by a red line connecting the two features between the two clones. The numbers on the ruler on top of the diagram for each clone corresponds to the DNA sequence in base pairs. Color-filled 'tracks' below the ruler show all annotated features for the clone, e.g. results from Blastx (BLX), tblastX (TBX), BlastN (BLN), Interpro HMMPfam (IHPF) and InterproScan (IPS). GeneScan (GSC) predicted exons can be found in the bottom track.

Creating a custom account

Users may create an account in PFGD by filling out some basic contact information. Registered users have the ability to save sequences and queries of interest to the MyData component of PFGD, as well download data in a variety of formats. In the future, registered users may also elect to receive emails regarding news, events and updates to the PFGD system. MyData is a tool for archiving and saving sequences and queries using personalized folder and query names. The MyData page supports bulk downloads of sequences in forward, reverse complement or both orientations in FASTA format as well as downloading sequence analysis and annotation details via the Summary Download option. Using the Summary Download feature, users may choose to download analysis results, EST quality screen details, cluster information, as well as feature location information for both genomic and transcript data in tab-delimited text or Excel formats. Access to user data is password protected: a temporary password is issued via email at the time of registration and can be changed when the user logs in. For the un-registered user, the interface supports individual sequence downloads in FASTA format.

Sequence comparisons

Users have the ability to BLAST (6) local sequences against PFGD EST, consensus, UniORF and genomic sequence data by pasting one or more FASTA formatted sequences into a conventional BLAST interface or by browsing the user's local hard drive for sequence files. Target database options include individual species for both genomic and transcript as well as EST, Consensus and UniORF datasets.

The comparative functional genomics browser (CFGB)

The CFGB visualizes genomic analysis results, including comparative transcript data aligned to genomic contigs to validate gene predictions (Figure 3). The CFGB also supports serial addition of contigs to the browser for comparative alignments and supports novel dynamic functional annotation comparisons by using description and/or GO term matches of the different analysis types. Each genomic sequence has

been annotated using the XGI genomic pipeline and each colored block represents the analysis type as well as the location and directionality of a match in relation to the genomic sequence. The size and orientation of the images in the CFGB can be manipulated by the zooming, panning and sorting functions. The ability to change the aspect ratios for all sequences at the same time is also supported.

Functional assay curation and display interface

PFGD supports curation and visualization of functional assay experiments. For each UniORF, a link is provided to the functional assay page. Here, authorized users can create and edit functional assay experiments performed using the UniORF. All users can view details on the functional assays experiments for all UniORFs.

SolGD interface

The SolGD interface allows users to query on Solanaceae transcript data for various species, such as *Nicotiana benthamiana*, *Nicotiana tabacum*, *Nicotiana otophora*, *Lycopersicon esculentum* and *Solanum tuberosum*. Most of the interface functionality available in PFGD is also available here.

Gene expression interface

The gene expression interface is accessible from SolGD. The 'Gene Expression' tab on top of the search filter page gives access to the gene expression data. These data can be accessed by either browsing by study/experiments in the database or searching by gene name. Various results pages allow a user to view their search results, study and sample details, and download images, raw data or normalized data for hybridizations. Researchers can bidirectionally traverse the data from either interesting expression results to putative functional assignments, or query for sequences of interest based on functional assignments and explore their expression profiles across experiments.

EXAMPLE OF A PFGD APPLICATION

To illustrate the functionality of PFGD, we describe a simple but powerful example. Let us say a user is interested in identifying 'genes encoding secreted proteins that uniquely expressed in zoospores of *P.infestans*'. A list of candidates can be easily generated by performing the following query. Libraries: *P.infestans*/transcript/zoospores, purified; sequence type: transcript/consensus; library constraint: strict; only show results with: PEXFinder results. In the current version (October 2005), this search generates 110 matching sequences. The user can further explore any interesting annotation associated with these sequences as well as assess the assigned features using the F&As link. Finally, the selected sequences could serve as a basis for wet lab experiments to determine the extent to which they are specific to the zoospore stage.

FUTURE DEVELOPMENTS

In the next year of development, PFGD will incorporate all the expression and functional data for *P.infestans* that will be generated by OSU-OARDC. A researcher using PFGD could query for a list of host plant genes expressed by a particular effector gene in or a list of effector genes that caused significant change in expression levels for a given host plant gene. Tools for various kinds of gene expression analysis will be integrated into PFGD as well.

Recently, the NSF/USDA microbial sequencing program has awarded MIT and NCGR to sequence the genomes of the oomycetes *P.infestans*, *Phytophthora capsici* and *Hyaloperonospora parasitica*. The genomic sequences generated by these projects will be annotated using the XGI genomic pipeline and integrated into PFGD.

ACKNOWLEDGEMENTS

We thank the National Science Foundation for funding this project under Plant Genome Research Program Grant Number 0211659. Funding to pay the Open Access publication charges for this article was provided by NSF.

Conflict of interest statement. None declared.

REFERENCES

1. Kamoun,S. (2003) Molecular genetics of pathogenic oomycetes. *Eukar. Cell*, **2**, 191–199.
2. Huitema,E., Bos,J.I., Tian,M., Win,J., Waugh,M.E. and Kamoun,S. (2004) Linking sequence to phenotype in *Phytophthora*–plant interactions. *Trends Microbiol.*, **12**, 193–200.
3. Waugh,M., Hraber,P., Weller,J., Wu,Y., Chen,G., Inman,J. and Kiphart.D., Sobral,B. (2000) The *Phytophthora* genome initiative database: informatics and analysis for distributed pathogenomic research. *Nucleic Acids Res.*, **28**, 87–90.
4. Randall,T.A., Dwyer,R.A., Huitema,E., Beyer,K., Cvitanich,C., Kelkar,H., Fong,A.M., Gates,K., Roberts,S., Yatzkan,E. *et al.* (2005) Large-scale gene discovery in the oomycete *Phytophthora infestans* reveals likely components of phytopathogenicity shared with true fungi. *Mol. Plant–Microbe. Interact.*, **18**, 229–243.
5. Inman,J.T., Flores,H.R., May,G.D., Weller,J.W. and Bell,C.J. (2000) A high-throughput distributed sequence analysis and database system. *IBM Syst. J.*, **40**, 464–486.
6. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
7. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res.*, **32**, D23–D26.
8. Henikoff,S. and Henikoff,J.G. (1994) Protein family classification based on searching a database of blocks. *Genomics*, **19**, 97–107.
9. Henikoff,S., Henikoff,J.G. and Pietrokovski,S. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, **15**, 471–479.
10. Zdobnov E.M. and Apweiler R. (2001), InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
11. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
12. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
13. Torto,T.A., Li,S., Styer,A., Huitema,E., Testa,A., Gow,N.A., van West,P. and Kamoun,S. (2003) EST mining and functional expression assays identify extracellular effector proteins from the plant pathogen *Phytophthora*. *Genome Res.*, **13**, 1675–1685.
14. 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.
15. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.