

GabiPD: the GABI primary database—a plant integrative ‘omics’ database

Diego Mauricio Riaño-Pachón^{1,2}, Axel Nagel¹, Jost Neigenfind¹, Robert Wagner¹, Rico Basekow¹, Elke Weber¹, Bernd Mueller-Roeber², Svenja Diehl³ and Birgit Kersten^{1,2,*}

¹GabiPD team, Bioinformatics group, Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1,

²Department of Molecular Biology, University of Potsdam, Karl-Liebknecht-Strasse 24-25, Haus 20, 14476

Potsdam-Golm and ³Former RZPD German Resource Center for Genome Research GmbH, Berlin, Germany

Received August 6, 2008; Accepted September 9, 2008

ABSTRACT

The GABI Primary Database, GabiPD (<http://www.gabipd.org/>), was established in the frame of the German initiative for Genome Analysis of the Plant Biological System (GABI). The goal of GabiPD is to collect, integrate, analyze and visualize primary information from GABI projects. GabiPD constitutes a repository and analysis platform for a wide array of heterogeneous data from high-throughput experiments in several plant species. Data from different ‘omics’ fronts are incorporated (i.e. genomics, transcriptomics, proteomics and metabolomics), originating from 14 different model or crop species. We have developed the concept of GreenCards for text-based retrieval of all data types in GabiPD (e.g. clones, genes, mutant lines). All data types point to a central Gene GreenCard, where gene information is integrated from genome projects or NCBI UniGene sets. The centralized Gene GreenCard allows visualizing ESTs aligned to annotated transcripts as well as displaying identified protein domains and gene structure. Moreover, GabiPD makes available interactive genetic maps from potato and barley, and protein 2DE gels from *Arabidopsis thaliana* and *Brassica napus*. Gene expression and metabolic-profiling data can be visualized through MapManWeb. By the integration of complex data in a framework of existing knowledge, GabiPD provides new insights and allows for new interpretations of the data.

INTRODUCTION

Experimental studies in the post-genomic era generate a very large amount of data from high-throughput experiments on biological systems. Current studies include,

among others, expression and metabolite profiles, proteome and interaction data (e.g. DNA–protein and protein–protein interactions), collected at different space and time scales. This increasing flow of data requires computational systems that, besides managing efficiently the enormous quantity of data, are capable of integrating and displaying these disparate data collections in a meaningful and user-friendly way. We have developed the GABI primary database, GabiPD, in order to fulfill these requirements.

GabiPD is a web-accessible database that was developed in the frame of the German initiative for Genome Analysis of the Plant Biological System (Genomanalyse im biologischen System Pflanze, GABI). GabiPD allows a seamless integration of varied ‘omics’ data types obtained from plant systems and will follow the MIAME (1) and MIAMET (2) standards for storing gene expression and metabolic-profiling data, respectively. Its flexible design allows for a high level of data integration, and eases cross-referencing the different GabiPD data types among each other (e.g. mapping information, sequences and single nucleotide polymorphisms (SNP), 2DE gel images and protein information) and access to public gene/protein-specific information, which in turn provides the users a comprehensive overview of the available information for their particular gene or protein of interest. The integration with genome databases like TAIR (3) and general nucleotide databases like GenBank, as well as cross-links to secondary databases, such as ARAMEMNON (4), PlnTFDB (5), GABI-KAT (6), PhosPhAt (7) and ProMEX (8) further increase the usefulness of GabiPD.

METHODS AND CONTENTS

Design and implementation

GabiPD’s web interface was developed using Perl and Java in combination with template processing to separate the visualization from the application logic. Our

*To whom correspondence should be addressed. Tel: +49-(0)331-567-8750; Fax: +49-(0)331-567-898-750; Email: kersten@mpimp-golm.mpg.de

applications are database driven, which means that the application interface logic is derived directly from the database structure (shown in blue in Figure 1). To achieve this, we deploy reverse engineering methods in combination with template processing to generate interfaces to programming languages like Perl or Java, thus supporting all the database-specific actions like ‘insert’, ‘update’, ‘delete’ or ‘select’. These object-oriented interfaces are automatically generated from the database schema, supporting inheritance, automated key generation and advanced exception handling. The separation of the database application interface from the application and visualization logic (shown in yellow in Figure 1) facilitates fast adjustment to modifications of the data structure and diminishes the efforts on fixing existing application logic during larger database changes.

GabiPD content and gene-centric views

Currently, GabiPD includes data originating from 14 different angiosperm species representing the most important lineages in the flowering plants (Figure 2). *Arabidopsis thaliana* is the most widely represented model species, followed by the crop plants *Solanum tuberosum* (potato) and *Hordeum vulgare* (barley). In GabiPD, genomic, transcriptomic, proteomic and metabolomic data are integrated from those species. Genomic data comprise mapping information, sequences and SNP/InDel information. Transcriptomics is represented by a large number of ESTs and corresponding sequence trace files. ESTs are further analysed by BLAST and ORF analysis. For barley, in addition, EST clustering results and corresponding information on a new 27K unigene set is accessible and downloadable. As a type of proteomic data, annotated 2DE gel images from *Arabidopsis thaliana* and *Brassica napus* are integrated. Moreover, transcript and metabolite-profiling data are provided via MapManWeb.

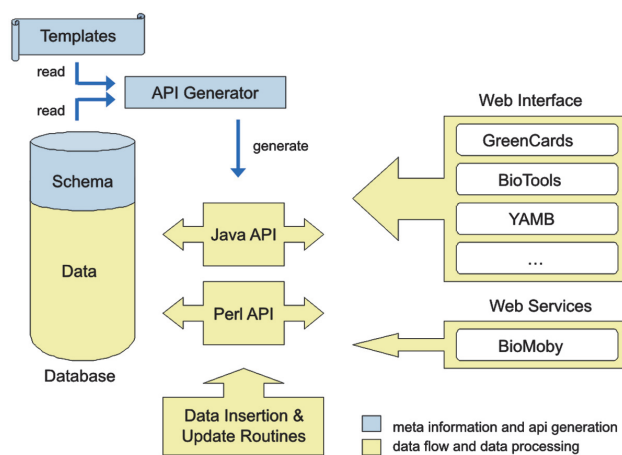


Figure 1. Schematic overview of the GabiPD structure. The ‘API Generator’ translates the ‘Templates’ using the database meta-information (shown in blue), generating the database application interface (Java and Perl API). The main applications, i.e. ‘Web Interface’, ‘Web Services’ and data manipulation routines, interacts with the ‘Database’ through the database application interface (shown in yellow).

Most entries of all GabiPD data types are pointing to the central Gene GreenCard and vice versa (see Figure 3 and next section). In the Gene GreenCard, gene information from genome annotation projects or NCBI UniGene sets is integrated and useful links to secondary databases are provided. Currently, the genome annotation (TAIR version 7.0) for *A. thaliana* (3) is integrated. Annotations for other sequenced species will follow. In order to ease the transfer of knowledge from sequenced to non-sequenced species, i.e. crop plants, we have performed similarity-based mappings between closely related species, i.e. *Arabidopsis* and *Brassica* spp.

Querying the database

We have developed the concept of ‘GreenCards’ as a central entry-point for text-based data queries and visualization, which grant public as well as credentials-based access to the integrated data in GabiPD. ‘GreenCards’ enable users to comprehensively query GabiPD by genotype name, marker or gene name, keyword or GenBank sequence accession number. Searches can be restricted to selected species or data types, while wildcards can be used to broaden the scope of the query. The result of a ‘GreenCard’ search is presented as a list of hits with links to complete descriptions, i.e. GreenCards. Figure 3 shows an example of this type of search, where the user had entered the gene name ‘FLOWERING LOCUS T’ as a search term. This search retrieves, among others, an

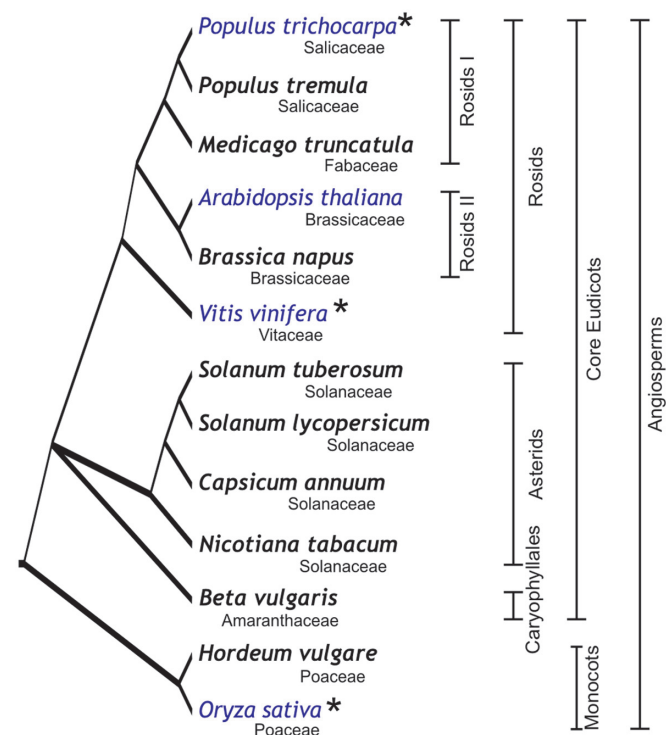


Figure 2. Phylogenetic tree depicting the evolutionary relationships among the species represented in GabiPD (15–17). Species for which whole-genome sequences and annotations are available are shown in blue. Asterisk indicates species that will soon be integrated in GabiPD. Species not shown: *S. bulbocastanum*, *S. demissum*, *S. phureja* and *S. spetzianii*.

Arabidopsis 'Gene GreenCard' (gene: AT1G65480.1) corresponding to the genome annotation project and 'Plant GreenCards' representing T-DNA insertion lines, e.g. plant 290E08, with flanking sequences that have BLAST hits to the gene AT1G65480.1 and with seeds available from GABI-KAT (6). Moreover, several 'Clone GreenCards' of cDNA clones (e.g. clone: MPMGp2011E01215) in which the keyword is found were retrieved by the search. A more strict relationship between the Gene GreenCard and the Plant and Clone GreenCards is established by similarity-based searches. The best BLAST hit of the sequence, e.g. representing relationship to a cDNA clone or a mutant plant line, appears

in the section 'Related with', and the users can go from the clone or the plant line to the associated gene or vice versa.

Furthermore, the 'Gene GreenCard', which displays information from genome annotation projects and NCBI UniGene sets, has been extended to include links to secondary databases, such as ARAMEMNON (4), GABI-KAT (6) and ProMEX (8). Additionally, schematic representations of gene sequence features are provided to highlight protein domains identified using the latest PFAM library (9), exon-exon borders and untranslated regions (UTRs) identified by the genome annotation projects (Figure 3). These features are displayed onto a representation of the cDNA sequence.

The figure shows a multi-panel screenshot of the GreenCards database interface.
 Panel A: Search results for 'FLOWERING LOCUS T' showing a list of genes and clones.
 Panel B: Detailed view of the Gene GreenCard for AT1G65480.1, including functional annotation (FT), related clones (MPMGp2011E01215), and external resources.
 Panel C: View of the clone MPMGp2011E01215, showing its sequence and related information.
 Panel D: A 2DE gel image with a protein spot highlighted by a blue cross-hair.
 Panel E: A BLAST alignment view showing sequence features and alignments.
 Panel F: A detailed protein description for FT, including its function, molecular weight, and sequence.
 Red arrows indicate the flow of information and user navigation between these panels.

Figure 3. Example of a keyword search using GreenCards. (A) The user had performed a search for the keywords 'FLOWERING LOCUS T', which retrieves links to the GreenCards of Genes (genome annotation projects), Clones (ESTs) and Plants (mutant plant lines). (B) Display of the Gene GreenCard, corresponding to the *Arabidopsis* annotated gene AT1G65480.1. Here, the users find high confidence matching EST sequences displayed in the 'Related with' section. Sequence features and the sequences themselves are displayed as well. The selected gene has a matching EST (Clone: MPMGp2011E01215), this Clone GreenCard is shown in (C) and links back to the Gene GreenCard and to the original EST trace file, displayed by JTree (D). A protein spot in rapeseed (*B. napus*) has been identified by 2DE/MS as the protein encoded by the retrieved gene, and a link directs the user to the 2DE gel image with the identified protein spot highlighted with blue cross-hair (E). The spot identified links to a description of the protein (F) that provides links to the original Gene GreenCard.

Alternatively, users can enter their own amino acid or nucleotide sequence to identify, by a BLAST search (10), similar sequences integrated in GabiPD.

In addition to the GreenCard and BLAST search functionality, users can browse and search the genetic maps and 2DE gels stored in GabiPD, through specifically designed visualization tools: (i) 2DEGelViewer by which

2DE gel images can be viewed in an interactive way, which allows retrieving extra information on 2D spots as identified by mass spectrometry (Figure 3); (ii) genetic mapping data can be visualized using YAMB (Yet Another Map Browser; Figure 4) with the possibility to view details on all mapped elements (11), (iii) MapManWeb (Figure 5) allows the visualization

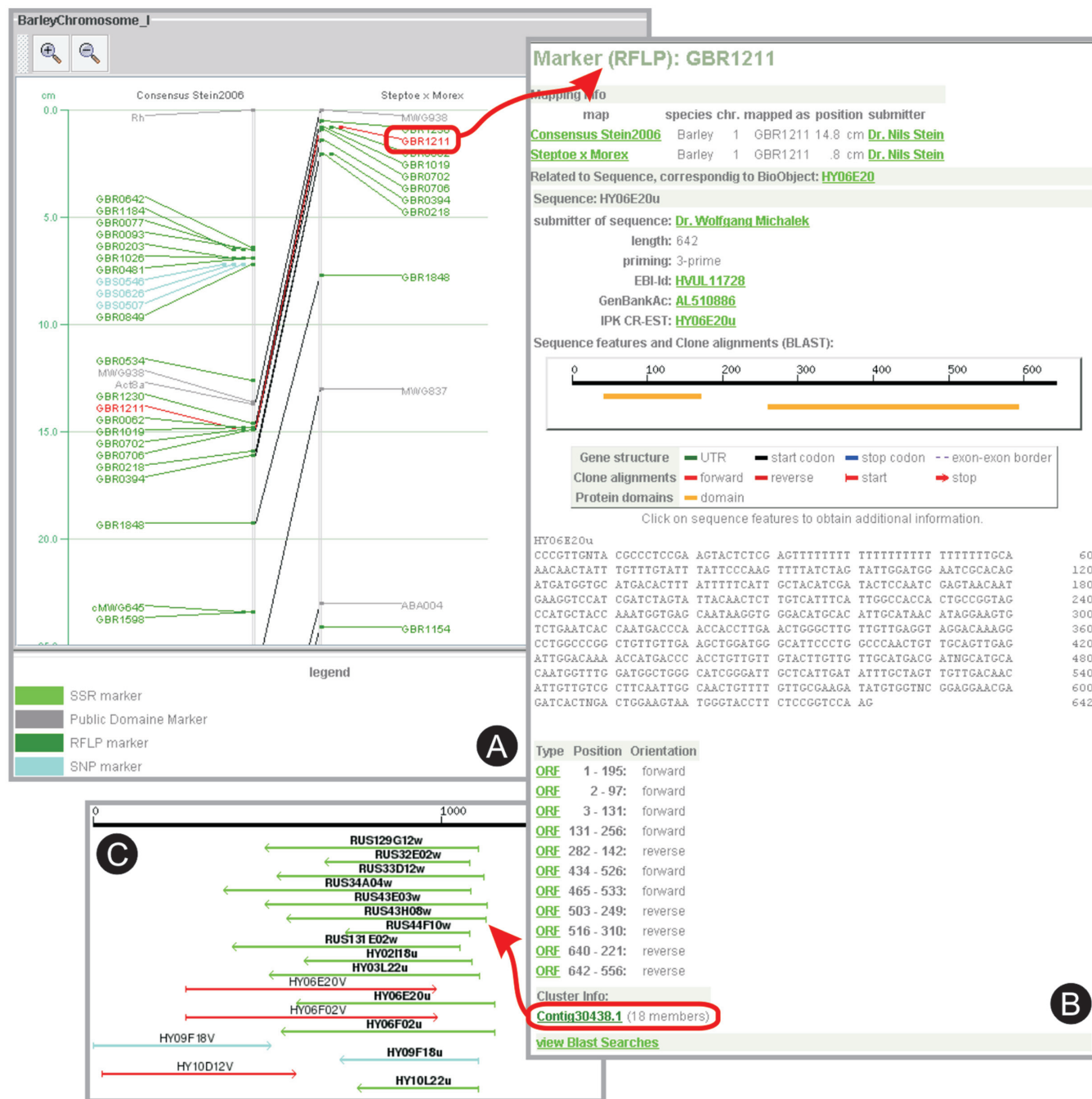


Figure 4. Visualization of the genetic maps published by Stein *et al.* (18). (A) The region between 0cM and 25 cM of barley chromosome I is shown in YAMB; markers with SNPs are shown in light blue, markers with restriction fragment length polymorphisms (RFLPs) are shown in dark green. A selected marker is displayed in red, and links to the Marker GreenCard (B), which contains information on a related EST therewith connecting genomic with transcriptomic information. With the EST description, cluster information is included (Contig30438.1) that links to the schematic representation of all ClusterContig members displayed onto the related consensus sequence (C). ESTs that were selected from this ClusterContig as representatives for the new 27K barley unigene set are shown in turquoise.

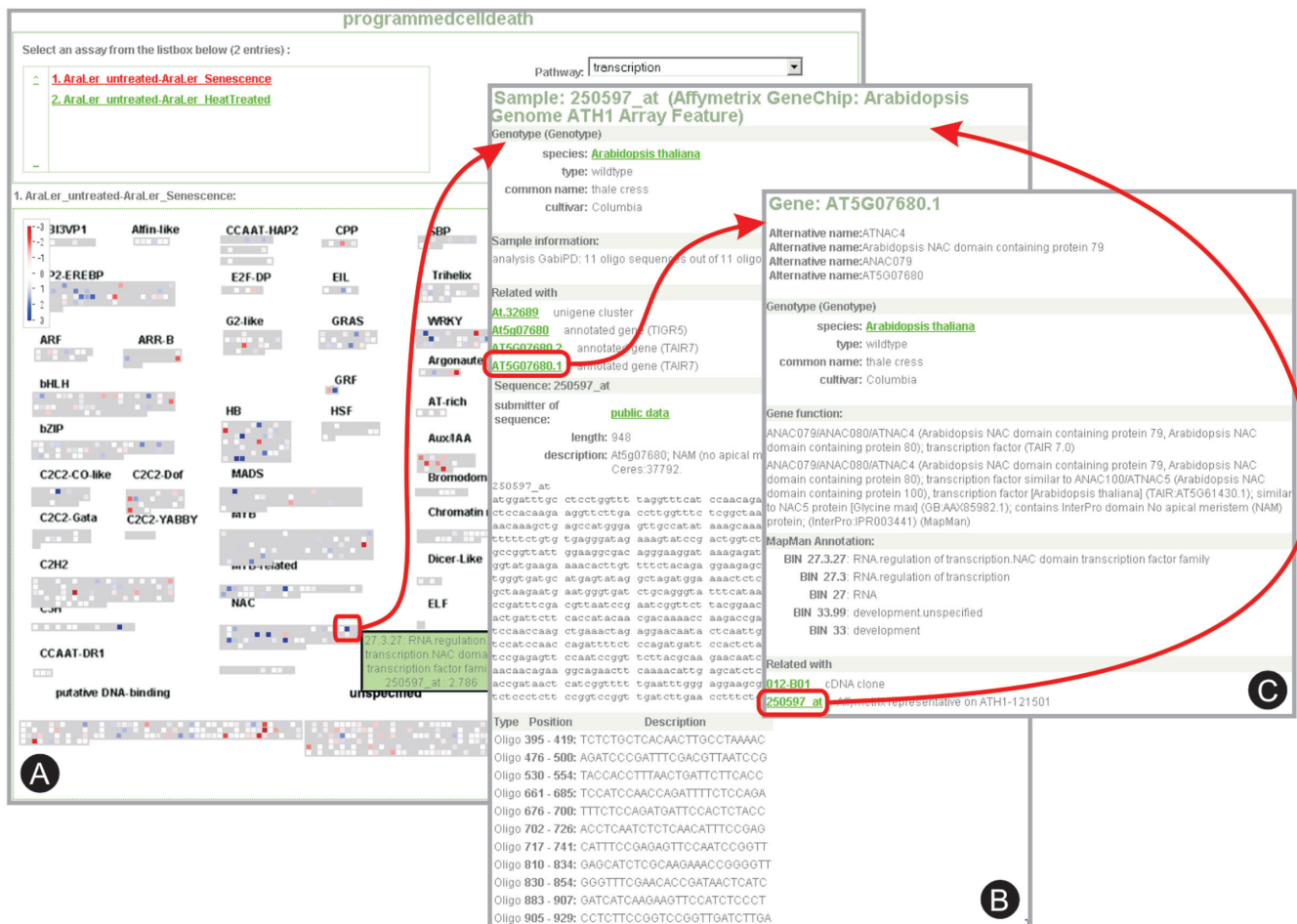


Figure 5. Visualization of expression profile data in MapManWeb. (A) The Affymetrix® NASC Array experiment on programmed cell death in *Arabidopsis* is displayed (NASCArrays reference number: 30). MapManWeb allows the visualization of expressed genes in different biological processes; here only probesets (i.e. genes) involved in transcription regulation are shown. (B) Details for a strongly down-regulated probeset, with links to the related Gene GreenCard in *A. thaliana*. (C) The Gene GreenCard for the selected gene (ANAC79) links back to the probeset of the Affymetrix® ATH1 array.

and extraction of relevant information from transcript and metabolite-profiling data and the graphical mapping of such data onto diagrams of metabolic pathways and other biological processes (12); and (iv) an extended version of JTrev (13) allows the display of sequence traces with integrated SNP information.

The GabiPD project page serves as an additional gateway to specific data by providing project-specific views, such as BreedCAM or PoMaMo (11) where potato genomic data and Solanaceae function maps for pathogen resistance are accessible.

ADDITIONAL TOOLS AVAILABLE FROM GabiPD

In addition to the data and data visualization available from our site, the newest versions of the following tools are made available for download:

MapMan desktop version: a user-driven software tool that displays large datasets (e.g. gene expression data from *Arabidopsis* Affymetrix arrays) onto diagrams of biological processes, such as metabolic pathways (12).

SATlotyper: a software tool designed for inferring haplotypes and phased genotypes from unphased SNP data for polyploid and polyallelic heterozygous populations (14).

FUTURE DIRECTIONS

The presentation of a wide spectrum of different plant species in GabiPD paves the way for cross-species comparisons that are facilitated by the availability of BLAST hits between the GabiPD sequences and plant NCBI UniGene sets. To ease the transfer of knowledge from sequenced to non-sequenced plant species, the genome annotations of *Oryza sativa*, *Populus trichocarpa* and *Vitis vinifera* will be added and mapped to closely related species in the near future. Furthermore, information about orthologous genes will be included for cross-species studies. Moreover, we will extend our WebServices to provide programmatic access to multiple data types for all plant species in GabiPD.

ACKNOWLEDGEMENTS

The GabiPD team thanks the GABI and the WPG (Wirtschaftsverbund Pflanzengenomforschung GABI e.V.) community for providing data and supporting the continuous development of the database. Dr Björn Usadel is acknowledged for helpful discussions and his support in the development of MapManWeb. We wish to thank Dr Patrick Schweitzer, Dr Lothar Altschmied, Dr Uwe Scholz and Dr Nils Stein for the collaboration in the generation of the new 27K barley unigene set and Dr Kathryn F. Beal for sharing the source code of the trace viewer JTrev. Özgür Demir and Sebastian Köhler are acknowledged for further developments of user interfaces and for data integration.

FUNDING

German Ministry for Education and Research (BMBF) (GABI I: 0312272, GABI II: 0313112 and GABI-FUTURE: 0315046); the former German Resource Center for Genome Research (RZPD) GmbH; Max Planck Society. Funding for open access charge: Max Planck Society.

Conflict of interest statement. None declared.

REFERENCES

- Ball,C.A. and Brazma,A. (2006) MGED standards: work in progress. *Omics*, **10**, 138–144.
- Jenkins,H., Hardy,N., Beckmann,M., Draper,J., Smith,A.R., Taylor,J., Fiehn,O., Goodacre,R., Bino,R.J., Hall,R. *et al.* (2004) A proposed framework for the description of plant metabolomics experiments and their results. *Nat. Biotechnol.*, **22**, 1601–1606.
- Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.
- Schwacke,R., Schneider,A., van der Graaff,E., Fischer,K., Catoni,E., Desimone,M., Frommer,W.B., Flugge,U.I. and Kunze,R. (2003) ARAMEMNON, a novel database for Arabidopsis integral membrane proteins. *Plant Physiol.*, **131**, 16–26.
- Riaño-Pachón,D.M., Ruzicic,S., Dreyer,I. and Mueller-Roeber,B. (2007) PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics*, **8**, 42.
- Li,Y., Rosso,M.G., Viehoveer,P. and Weisshaar,B. (2007) GABI-Kat SimpleSearch: an *Arabidopsis thaliana* T-DNA mutant database with detailed information for confirmed insertions. *Nucleic Acids Res.*, **35**, D874–D878.
- Heazlewood,J.L., Durek,P., Hummel,J., Selbig,J., Weckwerth,W., Walther,D. and Schulze,W.X. (2008) PhosPhAT: a database of phosphorylation sites in *Arabidopsis thaliana* and a plant-specific phosphorylation site predictor. *Nucleic Acids Res.*, **36**, D1015–D1021.
- Hummel,J., Niemann,M., Wienkoop,S., Schulze,W., Steinhäuser,D., Selbig,J., Walther,D. and Weckwerth,W. (2007) ProMEX: a mass spectral reference database for proteins and protein phosphorylation sites. *BMC Bioinformatics*, **8**, 216.
- Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
- 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.
- Meyer,S., Nagel,A. and Gebhardt,C. (2005) PoMaMo—a comprehensive database for potato genome data. *Nucleic Acids Res.*, **33**, D666–D670.
- Usadel,B., Nagel,A., Thimm,O., Redestig,H., Blaesing,O.E., Palacios-Rojas,N., Selbig,J., Hannemann,J., Piques,M.C., Steinhäuser,D. *et al.* (2005) Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses. *Plant Physiol.*, **138**, 1195–1204.
- Bonfield,J.K., Beal,K.F., Betts,M.J. and Staden,R. (2002) Trev: a DNA trace editor and viewer. *Bioinformatics*, **18**, 194–195.
- Neigenfind,J., Gyetvai,G., Basekow,R., Diehl,S., Achenbach,U., Gebhardt,C., Selbig,J. and Kersten,B. (2008) Haplotype inference from unphased SNP data in heterozygous polyploids based on SAT. *BMC Genomics*, **9**, 356.
- Soltis,P.S. and Soltis,D.E. (2004) The origin and diversification of angiosperms. *Am. J. Bot.*, **91**, 1614–1626.
- Angiosperm Phylogeny Group. (2003) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Bot. J. Linn. Soc.*, **141**, 399–436.
- Knapp,S. (2002) Tobacco to tomatoes: a phylogenetic perspective on fruit diversity in the Solanaceae. *J. Exp. Bot.*, **53**, 2001–2022.
- Stein,N., Prasad,M., Scholz,U., Thiel,T., Zhang,H., Wolf,M., Kota,R., Varshney,R.K., Perovic,D., Grosse,I. *et al.* (2007) A 1,000-loci transcript map of the barley genome: new anchoring points for integrative grass genomics. *Theor. Appl. Genet.*, **114**, 823–839.