D1144–D1151 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 29 November 2012*
doi:10.1093/nar/gks1153

MIPS PlantsDB: a database framework for comparative plant genome research

Thomas Nussbaumer, Mihaela M. Martis, Stephan K. Roessner, Matthias Pfeifer, Kai C. Bader, Sapna Sharma, Heidrun Gundlach and Manuel Spannagl*

Munich Information Center for Protein Sequences/Institute of Bioinformatics and Systems Biology, Helmholtz Center Munich—German Research Center for Environmental Health, 85764 Neuherberg, Germany

Received September 15, 2012; Revised October 24, 2012; Accepted October 25, 2012

ABSTRACT

The rapidly increasing amount of plant genome (sequence) data enables powerful comparative analyses and integrative approaches and also requires structured and comprehensive information resources. Databases are needed for both model and crop plant organisms and both intuitive search/browse views and comparative genomics tools should communicate the data to researchers and help them interpret it. MIPS PlantsDB (http://mips.helmholtz-muenchen.de/plant/genomes.jsp) was initially described in NAR in 2007 [Spannagl,M., Noubibou,O., Haase,D., Yang,L., Gundlach,H., Hindemitt, T., Klee,K., Haberer,G., Schoof,H. and Mayer,K.F. (2007) MIPSPlantsDB–plant database resource for integrative and comparative plant genome research. *Nucleic Acids Res.*, 35, D834–D840] and was set up from the start to provide data and information resources for individual plant species as well as a framework for integrative and comparative plant genome research. PlantsDB comprises database instances for tomato, *Medicago*, *Arabidopsis*, *Brachypodium*, *Sorghum*, maize, rice, barley and wheat. Building up on that, state-of-the-art comparative genomics tools such as CrowsNest are integrated to visualize and investigate syntenic relationships between monocot genomes. Results from novel genome analysis strategies targeting the complex and repetitive genomes of triticeae species (wheat and barley) are provided and cross-linked with model species. The MIPS Repeat Element Database (mips-REdat) and Catalog (mips-REcat) as well as tight connections to other databases, e.g. via web services, are further important components of PlantsDB.

INTRODUCTION

Sequencing of plant genomes has made a dramatic progress. With new high-volume data becoming available in ever shorter time periods, the challenges of data selection, integration, analysis and representation are the major driving forces for plant genome databases. The availability of plant genome sequence data from a wide range of taxa has shown to be extremely helpful in answering biological questions by comparative analyses (1,2). There is a need not only to store and provide both raw data and analyses results in informative and comprehensive structures but also to assist researches in exploring and analyzing data in up-to-date bioinformatic tools. MIPS PlantsDB is a plant database framework focusing on different core areas in plant genome research. Individual organism databases are provided for many important crop and model plants and new organisms and data are integrated in close collaboration with sequencing projects and plant infrastructure initiatives. Beyond individual genome databases, PlantsDB resources help to address specific questions in comparative and integrative plant genomics by providing tools to visualize synteny, transfer data from model systems to crops and explore similarities and peculiarities of different plant species. Repeat catalogs and classification systems for all plant species are further vital elements of PlantsDB.

MIPS PlantsDB is a member resource in the European Union-funded transPLANT initiative, a project aimed to enhance the inter-connectivity of distributed plant genome resources and databases within Europe and internationally and to support the development of plant genome resources as well as common standards. In this context, PlantsDB resources are complemented by BioMOBY (3) based web services that support seamless navigation and combination of services provided by PlantsDB and partner databases worldwide.

MIPS PlantsDB can be accessed at http://mips.helmholtz-muenchen.de/plant/genomes.jsp.

*To whom correspondence should be addressed. Tel: +49 89 3187 3581; Fax: +49 89 3187 3585; Email: manuel.spannagl@helmholtz-muenchen.de

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

Table 1. MIPS PlantsDB database instances and resources summary

MIPS PlantsDB instance	Species common name	Species scientific name	Genome sequence?	Gene annotations?	PlantsDB URL add-on	Release/version	Tools available
MAtDB	Thale cress	*Arabidopsis thaliana*	Yes	Yes	.../athal/index.jsp	TAIR10	GBr, OG
TomDB	Tomato	*Solanum lycopersicum*	Yes	Yes	.../tomato/index.jsp	ITAG V2.40	GBr, OG, CrN[b]
UrMeLDB	Barrel medic	*Medicago truncatula*	Yes	Yes	.../medi3/index.jsp	IMGAG Mt3.5v4	GBr, OG
MOsDB	Rice	*Oryza sativa*	Yes	Yes	.../rice/index.jsp	MSU6.1, RAP2	CrN, GBr, GZ, OG
MIPS Brachypodium DB	Purple false brome	*Brachypodium distachyon*	Yes	Yes	.../brachypodium/index.jsp	vl.2	CrN, GBr, GZ, OG
MIPS Sorghum DB	Sweet sorghum	*Sorghum bicolor*	Yes	Yes	.../sorghum/index.jsp	vl.4	CrN, GBr, GZ, OG
MIPS triticeae DB	Wheat, barley	*Triticum aestivum, Hordeum vulgare*	Yes[a]	Yes[a]	.../triticeae/index.jsp	misc.	CrN, GBr, GZ, OG
MGSP	Maize	*Zea mays*	Yes	Yes	.../maize/index.jsp	v5b.60	GBr, OG, CrN[b]

This table gives an overview about both species and genome resources stored in MIPS PlantsDB at this time.
[a]Wheat and barley genome sequence consists of sequence reads and partially assembled sequences. Gene annotations are also incomplete.
[b]In preparation.
Tools abbreviations: GBr, GBrowse instance available, sometimes hosted by external partner; CrN, species computed in CrowsNest; GZ, species included in GenomeZipper analyses; OG, pre-computed orthologous groups available.
MIPS PlantsDB URL: http://mips.helmholtz-muenchen.de/plant+add-on.

PLANTSDB—PLANT REFERENCE GENOME DATABASES

Starting with *Arabidopsis thaliana*, individual plant organism and reference databases have a long history in PlantsDB (4). Reference databases for both individual model and crop plant species provide researchers with high-quality structured and integrated data and support not only species-centric research but also facilitate comparative studies and knowledge transfer if embedded in a comparative analysis framework. MIPS PlantsDB hosts the European reference genome databases for the legume model organism *Medicago truncatula* (1), *Solanum lycopersicum* [tomato; (2)] and *Hordeum vulgare* [barley; (5)]. In close collaboration with international consortia and the European Union Framework 6/7 programs GrainLegumes, EUSOL and TriticeaeGenome, all genomic data generated are/were being integrated into the respective PlantsDB organism instances and presented to the research communities in structured formats and through different interfaces. In case of *Medicago*, tomato and barley, PlantsDB also serves as a central data integration hub for the structural and functional gene annotation with active and ongoing involvement in many aspects of the genome analysis, annotation curation and data management. PlantsDB also incorporates a number of additional important organism databases such as *A. thaliana*, *Oryza sativa* (rice), *Brachypodium distachyon*, *Sorghum bicolor* and *Zea mays* (maize). Data types available from the individual PlantsDB organism instances are summarized in Table 1.

MIPS PlantsDB individual organism databases are updated regularly, e.g. if new/updated external data releases become available or if substantial new data are generated or integrated through in-house analyses or within collaborations.

PLANTSDB—*TRITICEAE* INSTANCES

The family of *triticeae* plants includes many agronomical important species such as wheat, barley and rye but their genomes tend to be highly complex and repetitive. Wheat for instance is allo-hexaploid with a genome size of ~17 Gb and barley has a genome size of ~5.1 Gb (6).

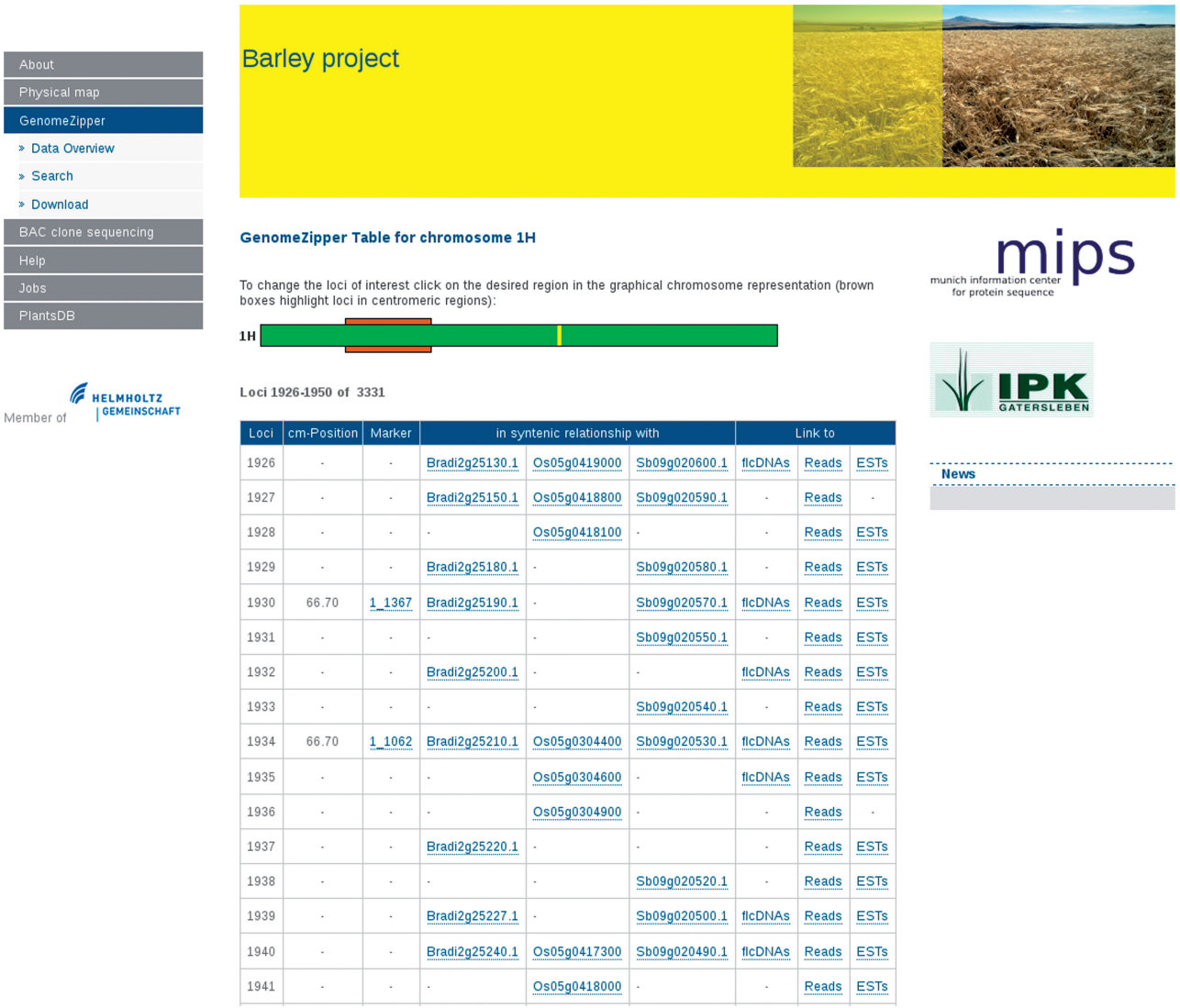
As a result, assemblies of whole-genome sequences of these important crop plants are extremely challenging even with improved algorithms and the latest sequencing technology. Lately, new strategies have been developed and applied to disclose the gene content of wheat and barley even in the presence of highly repetitive genome sequences (5,7).

For wheat, a 5× coverage 454 sequence of bread wheat (Chinese Spring line) was generated in UK (7). A Low Copy-number Genome assembly (LCG) was constructed by filtering out repetitive sequences and assembling the remaining low-copy sequences *de novo*. To avoid the collapsing of highly similar gene sequences from the three wheat sub-genomes, a set of orthologous representative grass genes incorporating genes from *B. distachyon*, *S. bicolor*, *O. sativa* and *H. vulgare* was generated in the first place.

Wheat raw reads were mapped and assembled on each Orthologous Group (OG) representative using stringent parameters, resulting in a large set of genic wheat sub-assemblies. Along with their linked OG representative, these genic wheat sub-assemblies will provide a helpful data foundation for researchers and breeders.

We therefore developed a wheat PlantsDB instance where both raw and processed data are available via intuitive search and download interfaces. http://mips.helmholtz-muenchen.de/plant/wheat/uk454survey/index.jsp gives access to homologous genic wheat sub-assembly sequences for any given gene identifier from the established grass reference organisms *Brachypodium*, *Sorghum*, rice and barley (as far as the gene is part of the initial clustering and associated with wheat sequences). A dedicated BLAST server allows users to search the orthologous representative grass gene set for any given query sequence. A FTP download server gives access to bulk download files.

A complimentary data resource is available from http://www.cerealsdb.uk.net/ (8) where e.g. SNPs derived from the UK wheat 454 sequences can be queried.



Loci	cm-Position	Marker	in syntenic relationship with			Link to		
1926	-	-	Bradi2g25130.1	Os05g0419000	Sb09g020600.1	flcDNAs	Reads	ESTs
1927	-	-	Bradi2g25150.1	Os05g0418800	Sb09g020590.1	-	Reads	-
1928	-	-	-	Os05g0418100	-	-	Reads	ESTs
1929	-	-	Bradi2g25180.1	-	Sb09g020580.1	-	Reads	ESTs
1930	66.70	1_1367	Bradi2g25190.1	-	Sb09g020570.1	flcDNAs	Reads	ESTs
1931	-	-	-	-	Sb09g020550.1	-	Reads	ESTs
1932	-	-	Bradi2g25200.1	-	-	flcDNAs	Reads	ESTs
1933	-	-	-	-	Sb09g020540.1	-	Reads	ESTs
1934	66.70	1_1062	Bradi2g25210.1	Os05g0304400	Sb09g020530.1	flcDNAs	Reads	ESTs
1935	-	-	-	Os05g0304600	-	flcDNAs	Reads	ESTs
1936	-	-	-	Os05g0304900	-	-	Reads	-
1937	-	-	Bradi2g25220.1	-	-	-	Reads	ESTs
1938	-	-	-	-	Sb09g020520.1	-	Reads	ESTs
1939	-	-	Bradi2g25227.1	-	Sb09g020500.1	flcDNAs	Reads	ESTs
1940	-	-	Bradi2g25240.1	Os05g0417300	Sb09g020490.1	flcDNAs	Reads	ESTs
1941	-	-	-	Os05g0418000	-	-	Reads	ESTs

Figure 1. Barley GenomeZipper in MIPS PlantsDB. This figure shows a region on barley chromosome 1H, constructed with the GenomeZipper concept. Detailed information and sequence download for anchored barley markers as well as 'zipped' reference organism genes, barley fl-cDNAs, ESTs and sequence reads can be obtained by clicking the individual links.

Although direct assembly of *triticeae* sequences is usually hampered by its repetitiveness, many grass genomes share a conserved gene order over large portions of their chromosomes (synteny) (9,10).

The GenomeZipper concept uses a novel approach that incorporates chromosome sorting, second-generation sequencing, array hybridization and systematic exploitation of conserved synteny with model grasses (11,12). Recently, this strategy allowed to assign 86% of the estimated 32 000 barley genes to individual chromosome arms. A series of bioinformatically constructed 'zippers' integrated gene indices of rice, *Sorghum* and *Brachypodium* in a conserved synteny model and assembled 21 766 barley genes in a putative linear order. As a result, the GenomeZipper provides an ordered, information-rich scaffold of the barley genome which can be queried by any anchored gene model from one of the grass model organisms (*Brachypodium*, *Sorghum* and rice). To assist this task, MIPS PlantsDB provides full access to both the barley GenomeZipper results and raw data through search and browse interfaces (http://mips.helmholtz-muenchen.de/plant/barley/gz/index.jsp). Figure 1 shows a screenshot of a specific chromosome region on the barley GenomeZipper.

In the lines of the barley zipper, a GenomeZipper for the wheat genome is currently constructed within IWGSC (http://www.wheatgenome.org/) and will be available with the same functionality soon.

Another *triticeae* data resource within PlantsDB complements the GenomeZipper for barley. We integrated several barley genetic maps with a physical map derived from high information content fingerprinting of 650 000 BAC sequences. A total of 570 000 (13× genome coverage) high-quality fingerprints were selected and entered the *de novo* contig assembly with FPC v9.0 (5). The resulting FPcontig map 'fpc_10' (9435 contigs,

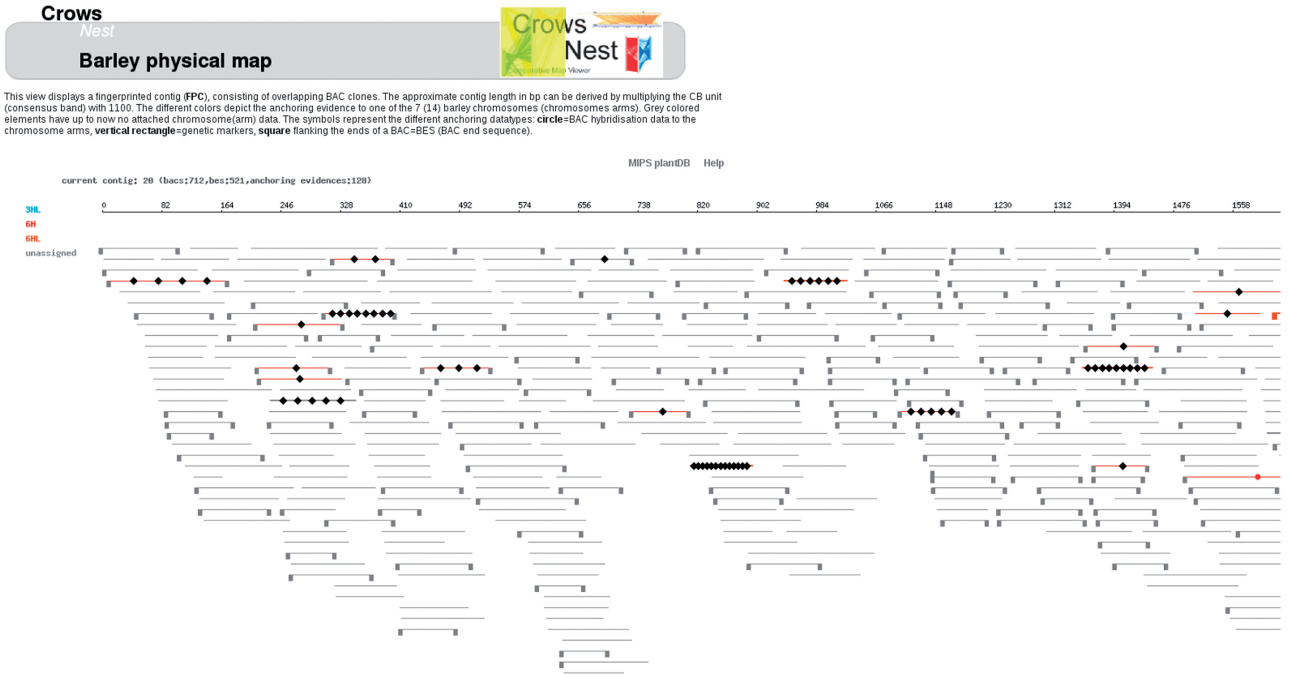


Figure 2. Visualization of the barley physical map in CrowsNest. A fingerprinted contig (FPC; example 'contig_20'), consisting of overlapping BAC clones. The different colors depict the anchoring evidence to one of the seven (14) barley chromosomes (chromosome arms). Gray-colored elements have no attached chromosome (arm) information. The symbols represent the different anchoring datatypes: circle = BAC hybridization data to the chromosome arms, vertical rectangle = genetic markers, square flanking the ends of a BAC = BES (BAC end sequence).

507 688 BAC clones) is currently displayed in a dedicated PlantsDB Gbrowse instance and positionally anchored FPcontigs are visualized within the CrowsNest tool (http://mips.helmholtz-muenchen.de/plant/barley/fpc/index.jsp). Figure 2 shows a screenshot of a barley FPcontig visualized within the CrowsNest tool.

PLANTSDB—COMPARATIVE GENOMICS TOOLS

Today more than ever the availability of intuitive comparative genome mapping and visualization tools is key to effectively address evolutionary questions, to transfer knowledge from a model plant to another genome of interest and to anchor unfinished genomes in the course of assembly.

Many tasks are automated processes in this area, but some steps still require human interpretation and direction. Visualization of inter- and intra-genome relationships, often over multiple scales, is critically important for these goals and also a key challenge because of the difficulty of the graphical representation. In addition, diverse sources and different techniques of generated genomics data need to be taken into account. For these reasons, a variety of highly specialized visualization tools have been developed so far (13–21). Some of these tools focus more on whole-genome alignment and synteny and others more on comparative analysis data displayed in feature tracks using either web-based or stand-alone approaches.

In recent years, a variety of strategies have been explored for graphically representing synteny at a whole-genome scale as well as at the chromosome level. Whole-genome views are usually graphically depicted using one of the following methods: (i) the historically well-known 2D 'dot plot', (ii) pill-shaped ideograms of reference genome chromosomes, banded and color-coded to indicate regions of synteny to a target genome and (iii) circular representation of reference genome chromosomes as arcs in a circle depicting aligned regions either as color-coded arcs in outer circles or as lines across the middle of the circle.

CrowsNest, a whole-genome interactive comparative mapping and visualization tool comparing genetic, physical and hierarchical (fingerprinted contigs) maps in the plant kingdom, builds on top of MIPS PlantsDB. The comparative map viewer is a web-based community resource and is integrated into the PlantsDB comparative genome framework. CrowsNest is specifically designed to visualize synteny at macro and micro levels. It allows to intuitively explore rearrangements, inversions, deletions at different resolutions, to transfer knowledge about function and conservation between several plant species and to derive evolutionary information.

CrowsNest can be accessed from the MIPS PlantsDB start page or directly at http://mips.helmholtz-muenchen.de/plant/crowsNest/index.jsp. Gene reports from included species provide direct links to the micro-syntenic views of the corresponding regions.

CrowsNest framework components, architecture and analysis pipeline

CrowsNest consists of two main integrated parts: (i) the web-based user interface with the integrated comparative visualization tool and (ii) the comparative analysis

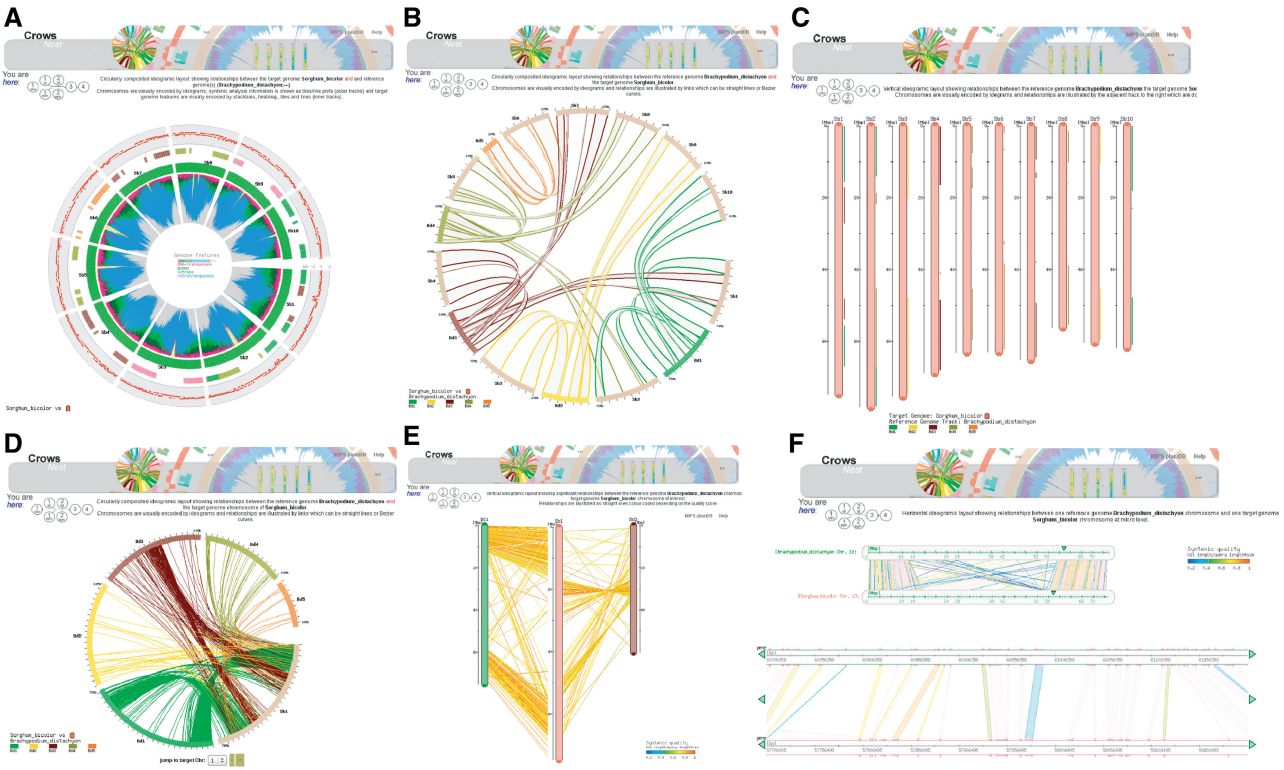


Figure 3. (**A**–**F**) CrowsNest visualization levels L1–L4. Different visualization levels of the CrowsNest tool for the reference grass organisms—*B. distachyon* and *S. bicolor*. The syntenic regions between the organisms can be browsed in a hierarchical way from macro-synteny (A–C) down to micro-synteny views (F). Navigation between the levels is possible by interactively selecting regions of interest in the views ('click zoom') or using the navigation bar.

pipeline. Putative orthologs were calculated using BLASTP between protein sequences of respective genomes. A sequence identity of at least 70% along with a hit length of at least 30 amino acids was required. The best bidirectional hits were extracted to avoid hits to paralogs and to cope with high sequence similarities among genes of the same gene family (22). For Level 1 and Level 3, gene pairs were grouped into syntenic segments. A sliding window approach with genome-specific window and shift sizes was applied (1 or 5 Mb for window size, 20% of window size used as shift). For paralogs, an increased sequence identity of >85% was used. All CrowsNest applications are implemented in OO Perl using wherever possible Bioperl modules for analysis purposes and the perl GD module for serving the map viewer with advanced graphics.

Comparative map viewer

The design of the comparative map viewer was driven by the idea to provide all three popular and well-established graphical representation methods to explore whole-genome relationships in the context of annotations and the alignment of unfinished and reference genomes. These three views are integrated into the top level view called L1. Altogether, the viewer consists of four levels (L1–L4) of different resolution ranging from whole-genome representations at L1 to specific region (<0.5 Mb) representations at L4. The navigational design is based on a top–down approach. Exploration is usually started at genome level view L1 with the option to 'drill down' from this macroscopic view to L2, the chromosome to whole-genome view, to L3, the chromosome to chromosome view, and then to L4. The graphical representation changes between L3 and L4 from a pill-shaped vertical to a pill-shaped horizontal one. In L4, the chromosomes being compared are 'stacked' on each other and with each chromosome the image map is extended vertically. The viewer has been designed to display a variety of features as tracks, such as the syntenic quality index, dN/dS ratio, repeat elements, gene family loci and others. Navigation between the levels is enabled as data are available. An example overview of the different view levels of CrowsNest is given in Figure 3.

CrowsNest currently harbors data from the model grass organisms—*B. distachyon*, *S. bicolor* and *O. sativa* (rice)—as well as from the crop plant *H. vulgare* (barley).

Figure 3A–E shows a whole-genome visualization of synteny depicting different levels of conservation. Presented are global orthologous relationships between the gene maps of *B. distachyon* and *S. bicolor* in Figure 3A–C. More detailed structural information together with conservation quality is illustrated in Figure 3D and E showing synteny to two reference chromosomes with the highest number of relationships to each of the target genome chromosomes. Switching to the pill-shaped

ideogram-like representation, regions of high conservation can clearly be seen. In this view, it can be switched between the orthologous view and the homology view, the latter indicating regions of duplication. The circular whole-genome view can also be used to illustrate synteny to multiple genomes.

Exploring conserved regions from macroscopic to microscopic views

All whole-genome views act as a starting point to explore synteny in a more depth investigation. Figure 3A–C illustrates the syntenic overview and let the investigator choose a chromosome of interest to navigate to the chromosome versus whole-genome view L2. Depending on the macro structure, a chromosome to chromosome relationship can be chosen to be directed to L3. By zooming, the magnification can be increased sufficiently high to display small-scale events of rearrangements, inversions and deletions. Breakpoints are derived easily at appropriate levels of resolution or can be compared to computational results from other sources. Below a resolution of 0.5 Mb, the level changes to L4 (example in Figure 3F) in which further zooming can be done and the elements are clickable for displaying feature/gene details stored in MIPS PlantsDB. Thus, CrowsNest allows seamless navigation and comparison from whole-genome views down to individual regions of interest and genetic elements located in these regions are directly linked to element entries and information in PlantsDB.

Exploring orthologous gene families—from model to crop genes

Complementing the more synteny-driven CrowsNest tool, MIPS PlantsDB also hosts a component for orthologous gene family construction and its comparative analysis. Orthologous gene families were computed for many PlantsDB species (including *Brachypodium*, rice, barley, *Sorghum*, maize and *Arabidopsis*) using OrthoMCL (23). In a first step, pairwise sequence similarities between all input protein sequences were calculated using BLASTP with an *e*-value cut-off of 1×10^{-5}. Markov clustering of the resulting similarity matrix was used to define the ortholog cluster structure, using an inflation value (−I) of 1.5 (OrthoMCL default). The results of this orthologous gene family constructions can be accessed from every individual gene family report (such as from the gene report of a specific *Brachypodium* or *Arabidopsis* gene of interest) where cross-references to all other genes in the same family are provided. Using this information (the gene identifiers in the orthologous gene family), corresponding orthologous genic sequences from bread wheat can be derived from PlantsDB using the procedure outlined in the *triticeae* section. The same workflow can also be applied other way around, e.g. starting from an unknown wheat sequence. This sequence can be searched against the wheat genic sequences and the OG representatives (see '*triticeae*' section for details), positively resulting in a grass reference gene model and its associated wheat orthologous genic sequences. This grass reference gene model can then be searched in PlantsDB and its gene report gives full access to the containing orthologous gene family and closely related grass genes.

MIPS REPEAT ELEMENT DATABASE (mips-REdat) AND CATALOG (mips-REcat)

Plant genomes are crowded by taxon-specific mobile elements and their deteriorated remnants, with portions between 20% and >90% of primarily LTR-retrotransposon insertions leading to complex and highly repetitive structures (24). Transposons play mostly harmful and sometimes long-term beneficial roles in evolutionary processes (25). The interplay between proliferation and removal of transposable elements greatly influences genome size and chromosomal architecture. Their prominent differential accumulations, even within closely related species, pose intriguing questions about host control, transposon countermeasures and the conditions disturbing the balance.

Our plant repeat database mips-REdat was set up in conjunction with mips-REcat, a detailed hierarchical repeat classification catalog to facilitate a consistent cross-species comparative transposon annotation. This resource is both useful for characterizing and comparing the transposon complements of different species or sequence sets as well as for repeat masking prior to gene annotation, to reduce computing time and to minimize unwanted transposon-related gene calls. Initially, mips-REdat contained a compilation of publicly available plant transposon sequences like Trep (http://wheat.pw.usda.gov/ITMI/Repeats/), TIGR repeats (26) or Repbase (27) and was rapidly filled up with up to now ~37 000 *de novo* detected LTR-retrotransposon and ~300 DNA transposon sequences from the genomes presented in MIPS PlantsDB. The REdat sequences are characterized by REcat keys (28), which in turn are mapped to the common transposon classifications of (29) and (30). The current public version mips-REdat_v9.0p consists of ~42 000 non-redundant sequences, which were clustered with ≥95% identity over ≥95% length coverage. They add up to ~350 Mb, stem from 44 species and cover 20 different genera. The public release (ftp://ftpmips.helmholtz-muenchen.de/plants/REdat/) does not contain yet unpublished data or sequences from Repbase and is subjected to regular updates. REdat can also be accessed on our website with the option to retrieve customized fasta files by repeat type and taxonomy.

PLANTSDB—transPLANT

transPLANT (Trans-national Infrastructure for Plant Genomic Science) is an EU project bringing together 11 institutions involved in plant data integration, management and analysis.

One of the missions of transPLANT is to provide a comprehensive set of computational and interactive services to the plant research community by developing distributed but tightly connected resources.

MIPS PlantsDB is part of that consortium and responsible for creating and maintaining a registry of important sequence-based resources for species of agricultural and economic importance as well as model systems.

We collected repository data for publicly available plant genome database systems maintained by both transPLANT and non-transPLANT partners.

A total of 187 distinct plant genome resources are registered at the transPLANT data registry at this time. The registry can be queried both at MIPS PlantsDB (http://mips.helmholtz-muenchen.de/plant/transplant/index.jsp) and at the official transPLANT web hub at EBI (http://transplantdb.eu/, synchronized with PlantsDB regularly) for e.g. keywords, species names and data types.

Changes and updates to the registry can be performed by database providers soon, lowering the maintaining cost and ensuring expert-curated and -driven information.

CONCLUSIONS

Since initially described in NAR in 2007 (28), MIPS PlantsDB was significantly extended both in plant genome data and retrieval and analysis functionality. The database framework integrates genome data from both model and crop plants and facilitates knowledge transfer between them using state-of-the-art comparative genomics tools such as CrowsNest and the GenomeZipper concept. MIPS PlantsDB is closely connected to the barley and wheat communities and provides access to the latest data generated within. Since much of these data are complex, intuitive and step-by-step interfaces and comparative genomics tools were developed and integrated. As data curation manpower is limited and thus plant genomic data resources risk to erode for individual data resources below a critical size, the transPLANT project provides infrastructure, knowledge and logical backbone to closely connect distributed plant genome resources in Europe and in conjunction with international partners.

ACKNOWLEDGEMENTS

The authors thank all our collaboration partners and data contributors. For barley these include IPK (Nils Stein, Uwe Scholz), FLI (Matthias Platzer), JKI (Frank Ordon), James Hutton Institute (Robbie Waugh), IEB (Jaroslav Doležel), University of Udine (Michele Morgante), UCR (Tim Close), ACPFG (Peter Langridge) and many more. For wheat these include: Centre for Genome Research, University of Liverpool (Neil Hall, Anthony Hall, Rachel Brenchley), School of Biological Sciences, University of Bristol (Keith J. Edwards, Gary L.A. Barker), John Innes Centre (Michael W. Bevan) as well as IWGSC (International Wheat Genome Sequencing Consortium, Kellye Eversole, Catherine Feuillet, INRA, Jane Rogers, Jon Wright and Mario Caccamo, TGAC, Norwich, UK) and many more. The authors also thank SGN (sol genomics network, Lukas Mueller) and ITAG (International Tomato Annotation Group) for collaborating in the tomato project as well as JCVI (Chris Town), Nevin Young and IMGAG (International Medicago Genome Annotation Group) for collaborating in the Medicago genome project.

FUNDING

The European Commission (framework 6 programme) within the Grain Legumes Integrated Project (GLIP) and EU-SOL project; the framework 7 programme in the TriticeaeGenome project and transPLANT project (funded by the European Commission within its 7th Framework Programme under the thematic area 'Infrastructures', contract number 283496); the German Ministry for Education and Research (BMBF) within the GABI and Plant for the future projects TRITEX and BARLEX. Funding for open access charge: Helmholtz Association.

Conflict of interest statement. None declared.

REFERENCES

1. Young,N.D., Debelle,F., Oldroyd,G.E., Geurts,R., Cannon,S.B., Udvardi,M.K., Benedito,V.A., Mayer,K.F., Gouzy,J., Schoof,H. *et al*. (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, **480**, 520–524.
2. Tomato Genome,C. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.
3. Wilkinson,M., Schoof,H., Ernst,R. and Haase,D. (2005) BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. The PlaNet exemplar case. *Plant Physiology*, **138**, 5–17.
4. Schoof,H., Ernst,R., Nazarov,V., Pfeifer,L., Mewes,H.W. and Mayer,K.F. (2004) MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource for plant genomics. *Nucleic Acids Res.*, **32**, D373–D376.
5. Consortium,T.I.B.G.S. (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711–716.
6. Dolezel,J. and Bartos,J. (2005) Plant DNA flow cytometry and estimation of nuclear genome size. *Ann. Botany*, **95**, 99–110.
7. Brenchley,R., Pfeifer,M., Barker,G.L.A., D'Amore,R., Allen,A.M., McKenzie,N., Kramer,M., Kerhornou,A., Bolser,D., Kay,S. *et al*. (2012) Analysis of the bread wheat genome using whole genome shotgun sequencing. *Nature*.
8. Wilkinson,P.A., Winfield,M.O., Barker,G.L., Allen,A.M., Burridge,A., Coghill,J.A. and Edwards,K.J. (2012) CerealsDB 2.0: an integrated resource for plant breeders and scientists. *BMC Bioinformatics*, **13**, 219.
9. Moore,G., Devos,K.M., Wang,Z. and Gale,M.D. (1995) Cereal genome evolution. Grasses, line up and form a circle. *Curr. Biol.*, **5**, 737–739.
10. Devos,K.M. (2005) Updating the 'crop circle'. *Curr. Opin. Plant Biol.*, **8**, 155–162.
11. Mayer,K.F., Taudien,S., Martis,M., Simkova,H., Suchankova,P., Gundlach,H., Wicker,T., Petzold,A., Felder,M., Steuernagel,B. *et al*. (2009) Gene content and virtual gene order of barley chromosome 1 H. *Plant Physiol.*, **151**, 496–505.
12. Mayer,K.F., Martis,M., Hedley,P.E., Simkova,H., Liu,H., Morris,J.A., Steuernagel,B., Taudien,S., Roessner,S., Gundlach,H. *et al*. (2011) Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell*, **23**, 1249–1263.
13. Sinha,A.U. and Meller,J. (2007) Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics*, **8**, 82.
14. Lyons,E., Pedersen,B., Kane,J., Alam,M., Ming,R., Tang,H., Wang,X., Bowers,J., Paterson,A., Lisch,D. *et al*. (2008) Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.*, **148**, 1772–1781.

15. Krzywinski,M., Schein,J., Birol,I., Connors,J., Gascoyne,R., Horsman,D., Jones,S.J. and Marra,M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.
16. Carver,T., Berriman,M., Tivey,A., Patel,C., Bohme,U., Barrell,B.G., Parkhill,J. and Rajandream,M.A. (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, **24**, 2672–2676.
17. Youens-Clark,K., Faga,B., Yap,I.V., Stein,L. and Ware,D. (2009) CMap 1.01: a comparative mapping application for the Internet. *Bioinformatics*, **25**, 3040–3042.
18. Meyer,M., Munzner,T. and Pfister,H. (2009) MizBee: a multiscale synteny browser. *IEEE Transact. Visual. Comput. Graph.*, **15**, 897–904.
19. Brendel,V., Kurtz,S. and Pan,X. (2007) Visualization of syntenic relationships with SynBrowse. *Methods Mol. Biol.*, **396**, 153–163.
20. Tang,H., Wang,X., Bowers,J.E., Ming,R., Alam,M. and Paterson,A.H. (2008) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.*, **18**, 1944–1954.
21. Van Bel,M., Proost,S., Wischnitzki,E., Movahedi,S., Scheerlinck,C., Van de Peer,Y. and Vandepoele,K. (2012) Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiol.*, **158**, 590–600.
22. Martinez,M. (2011) Plant protein-coding gene families: emerging bioinformatics approaches. *Trends Plant Sci.*, **16**, 558–567.
23. Li,L., Stoeckert,C.J. Jr and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
24. Vitte,C. and Panaud,O. (2005) LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenet. Genome Res.*, **110**, 91–107.
25. Chenais,B., Caruso,A., Hiard,S. and Casse,N. (2012) The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. *Gene*, **509**, 7–15.
26. Ouyang,S. and Buell,C.R. (2004) The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res.*, **32**, D360–D363.
27. Jurka,J., Kapitonov,V.V., Pavlicek,A., Klonowski,P., Kohany,O. and Walichiewicz,J. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, **110**, 462–467.
28. Spannagl,M., Noubibou,O., Haase,D., Yang,L., Gundlach,H., Hindemitt,T., Klee,K., Haberer,G., Schoof,H. and Mayer,K.F. (2007) MIPSPlantsDB–plant database resource for integrative and comparative plant genome research. *Nucleic Acids Res.*, **35**, D834–D840.
29. Wicker,T., Sabot,F., Hua-Van,A., Bennetzen,J.L., Capy,P., Chalhoub,B., Flavell,A., Leroy,P., Morgante,M., Panaud,O. *et al.* (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*, **8**, 973–982.
30. Kapitonov,V.V. and Jurka,J. (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet.*, **9**, 411–412; author reply 414.