HAL
open science

TransPLANT resources for triticeae genomic data

Manuel Spannagl, Michael M. Alaux, Matthias Lange, Dan M. Bolser, Kai Christian Bader, Thomas Letellier, Erik Kimmel, Raphaël-Gauthier R.-G. Flores, Cyril Pommier, Arnaud Kerhornou, et al.

To cite this version:

Manuel Spannagl, Michael M. Alaux, Matthias Lange, Dan M. Bolser, Kai Christian Bader, et al.. TransPLANT resources for triticeae genomic data. PLANT GENOME, 2016, 9 (1), 13 p. 10.3835/plantgenome2015.06.0038 . hal-02635899

HAL Id: hal-02635899

https://hal.inrae.fr/hal-02635899v1

Submitted on 27 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Published February 26, 2016

ORIGINAL RESEARCH

transPLANT Resources for Triticeae Genomic Data

Manuel Spannagl,* Michael Alaux, Matthias Lange, Daniel M. Bolser, Kai C. Bader, Thomas Letellier, Erik Kimmel, Raphael Flores, Cyril Pommier, Arnaud Kerhornou, Brandon Walts, Thomas Nussbaumer, Christoph Grabmuller, Jinbo Chen, Christian Colmsee, Sebastian Beier, Martin Mascher, Thomas Schmutzer, Daniel Arend, Anil Thanki, Ricardo Ramirez-Gonzalez, Martin Ayling, Sarah Ayling, Mario Caccamo, Klaus F.X. Mayer, Uwe Scholz, Delphine Steinbach, Hadi Quesneville, and Paul J. Kersey

Abstract

The genome sequences of many important Triticeae species, including bread wheat (*Triticum aestivum* L.) and barley (*Hordeum vulgare* L.), remained uncharacterized for a long time because their high repeat content, large sizes, and polyploidy. As a result of improvements in sequencing technologies and novel analyses strategies, several of these have recently been deciphered. These efforts have generated new insights into Triticeae biology and genome organization and have important implications for downstream usage by breeders, experimental biologists, and comparative genomicists. transPLANT (http://www.transplantdb.eu) is an EU-funded project aimed at constructing hardware, software, and data infrastructure for genome-scale research in the life sciences. Since the Triticeae data are intrinsically complex, heterogenous, and distributed, the transPLANT consortium has undertaken efforts to develop common data formats and tools that enable the exchange and integration of data from distributed resources. Here we present an overview of the individual Triticeae genome resources hosted by transPLANT partners, introduce the objectives of transPLANT, and outline common developments and interfaces supporting integrated data access.

Published in The Plant Genome 9
doi: 10.3835/plantgenome2015.06.0038

© Crop Science Society of America
5585 Guilford Rd., Madison, WI 53711 USA
This is an open access article distributed under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

CROPS from the tribe of the Triticeae, including wheat, barley, and rye (*Secale cereale* L.), account for some of the most important nutritional resources in the human diet. Until recently, genomics-informed breeding approaches were limited in Triticeae species, as few genomic data were available. This lack can mainly be attributed to the inherent complexity of their genomes and genetics, especially in species of high economic interest such as barley and bread wheat. With estimated haploid and triploid sizes of 5.3 and 17.1 Gb, respectively, the genomes of these (and other Triticeae) species significantly exceed the size of the haploid human genome (~3 Gb). The high overall repeat content and, in bread wheat,

M. Spannagl, K.C. Bader, T. Nussbaumer, and K.F.X. Mayer, Plant Genome and Systems Biology (PGSB), Helmholtz Center Munich, D-85764, Neuherberg, Germany; M. Alaux, T. Letellier, E. Kimmel, R. Flores, C. Pommier, D. Steinbach, and H. Quesneville, INRA, UR1164 URGI–Research Unit in Genomics-Info, INRA de Versailles, Route de Saint-Cyr, Versailles, 78026, France; M. Lange, J. Chen, C. Colmsee, S. Beier, M. Mascher, T. Schmutzer, D. Arend, and U. Scholz, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Stadt Seeland, D-06466, Germany; D.M. Bolser, A. Kerhornou, B. Walts, C. Grabmuller, and P.J. Kersey, European Molecular Biology Lab., The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK; A. Thanki, R. Ramirez-Gonzalez, M. Ayling, S. Ayling, and M. Caccamo, The Genome Analysis Centre (TGAC), Norwich Research Park, Norwich, NR4 7UH, UK. Received 8 June 2015. Accepted 14 Oct. 2015. *Corresponding author (manuel.spannagl@helmholtz-muenchen.de).

Abbreviations: BAC, bacterial artificial chromosome; CSS, chromosome survey sequence; EBI, European Bioinformatics Institute; EST, expressed sequence tag; GO, gene ontology; IBSC, International Barley Genome Sequencing Consortium; IPK, Leibniz Institute of Plant Genetics and Crop Plant Research; IWGSC, International Wheat Genome Sequencing Consortium; PGSB, Plant Genome and Systems Biology unit at the Helmholtz Center Munich; QTL, quantitative trait loci; RNA-seq, RNA sequencing; SNP, single nucleotide polymorphism; TGAC, The Genome Analysis Centre; URGI, Unité de Recherche Génomique Info at the Institut National de la Recherche Agronomique; WheatIS, International Wheat Information System.

the allohexaploid genome structure, further complicate every aspect of the collation and analysis of genomic data, including sequencing, assembly, gene calling, and functional analysis.

Nevertheless, great progress has recently been made in deciphering the genome sequences and gene content of these species through coordinated international collaboration. The International Barley Genome Sequencing Consortium (IBSC) reported a draft genome sequence for the barley cultivar Morex obtained using mainly next-generation sequencing technologies, chromosome sorting, an integrated physical and genetic map, gene predictions, and analysis of the transcriptional landscape (IBSC, 2012). For bread wheat, 5× 454 whole-genome sequencing was used to generate assemblies of homeologous genes and to analyze the gene content of hexaploid wheat in a reference-directed approach (Brenchley et al., 2012). In 2014, the International Wheat Genome Sequencing Consortium (IWGSC) released draft genome sequence assemblies for all wheat chromosome arms (again, following the use of a physical sorting strategy before sequencing), with a total of 124,201 annotated gene models and supporting transcriptomics data (IWGSC, 2014). Meanwhile, a French-led consortium has generated a reference sequence and gene annotation for wheat chromosome 3B (Choulet et al., 2014). Genomic data for a number of additional Triticeae species has also been generated recently, including the bread wheat subgenome progenitors *Aegilops tauschii* Coss. (Jia et al., 2013), *A. sharonensis* Eig and *A. speltoides* Tausch (IWGSC, 2014), and *T. urartu* Tumanian ex Gandilyan (Ling et al., 2013), and the tetraploid *T. turgidum L. subsp. durum* (Desf.) Husn. (pasta wheat) (IWGSC, 2014). GenomeZippers, a synteny-enabled anchoring approach, have been constructed for the Triticeae species barley (Mayer et al., 2011), wheat (IWGSC, 2014), *A. tauschii* (Luo et al., 2013), and rye (Martis et al., 2013), facilitating the positioning of 10,000s of genes in the absence of finished genome sequences. While the sequence of all these species is incompletely assembled, significant progress has been made toward assembling the gene space, assigning contigs to chromosomes, and ordering them. The data is already sufficient to support analyses of gene families, variation within populations, large scale synteny, and association of genotype with phenotype.

The availability of genomic data from multiple Triticeae species is expected to facilitate powerful comparative genomics approaches and help to enhance understanding of Triticeae biology and evolution. However, the use of multiple approaches to genome sequencing and assembly (e.g., chromosome sorting, GenomeZippers, and genome survey sequencing), the variety of associated data types (e.g., gene predictions, expression data, and molecular markers), and the existence of alternative coordinate systems (e.g., genetic map, physical map, and numerical position in molecular sequence) can make it difficult for users to combine different data sets easily and correctly. Storage, integration, and visualization of these heterogenous and complex data are essential to enable efficient research.

Here we describe Triticeae genome data resources maintained by partners in the EU-funded transPLANT project. The transPLANT project aims at producing an integrated, coherent data infrastructure shared among dispersed expert resources with strong interconnections between them (including cross-linking and the use of common formats, tools, and datasets) designed to make it easy for users to switch between different resources according to which one best addresses their current point of interest.

Results

transPLANT

The transPLANT project (Table 1, reference no. 1 [Table 1.1]; hereafter, this format is used to reference Table 1) is an integrated infrastructure funded by the Framework 7 program of the European Union. It brings together 11 partners from seven countries with the aim of developing common standards, data, and technologies in the plant genomics area. A major focus of the work is variation data and the development of tools to organize, archive, and analyze this. Another major focus is the definition and use of common reference sequences so that annotation from different resources can be shared and compared. A third focus is the development of a distributed query infrastructure to provide a common point of entry to data held in multiple, dispersed resources.

In the context of Triticeae data, five transPLANT partners (The Plant Genome and Systems Biology unit at the Helmholtz Center Munich [PGSB], the European Bioinformatics Institute [EBI], the Unité de Recherche Génomique Info at the Institut National de la Recherche Agronomique [URGI], the Leibniz Institute of Plant Genetics and Crop Plant Research [IPK] and The Genome Analysis Centre [TGAC]) are involved and interacting with other partners engaged in the international consortia (IWGSC and IBSC) coordinating the sequencing and assembly of these genomes. Figure 1 provides an overview over the respective Triticeae data resources hosted by transPLANT partners together with available species and shared search interfaces and services. Data from barley, wheat, and other species have been exported to common data formats in accordance with established standards for data representation, exchanged between partners, and synchronized across the distributed transPLANT resources. Moreover, partners have collaborated in comparative analysis of Triticeae genomes to study ancient duplications, polyploidy, and syntenic relationships.

As a result, end users benefit from a number of specialized tools and interfaces operating on synchronized Triticeae genome data hosted and exchanged by the transPLANT partners. This includes an interface embedded in the transPLANT web portal enabling keyword searches over the data inventories of many transPLANT partners as well as extensive cross-linking between

Table 1. List of URLs for transPLANT Triticeae resources, tools, and websites. References to specific URLs are given in the format Table 1. REFERENCE no. (e.g. Table 1.1) throughout the text.

Reference no.	Service provider	URL	Short description
1	transPLANT	http://www.transplantdb.eu	transPLANT web hub
2	transPLANT	http://www.transplantdb.eu/resource-registry	transPLANT resource registry
3	PGSB	http://pgsb.helmholtz-muenchen.de/plant/transplant/genomeResources.jsp	transPLANT resource registry mirror at PGSB
4	PGSB	http://pgsb.helmholtz-muenchen.de/plant/triticeae/index.jsp	PlantsDB Triticeae homepage
5	PGSB	http://pgsb.helmholtz-muenchen.de/plant/crowsNest/index.jsp	PlantsDB CrowsNest tool
6	EBI	http://plants.ensembl.org	Ensembl Plants homepage
7	INRA	http://wheat-urgi.versailles.inra.fr/	INRA URGI wheat database
8	INRA	http://wheat-urgi.versailles.inra.fr/Seq-Repository	IWGSC wheat sequence repository
9	INRA	https://urgi.versailles.inra.fr/gb2/gbrowse/wheat_phys_pub	Wheat physical maps browser
10	INRA	https://urgi.versailles.inra.fr/blast/	URGI wheat BLAST search tool
11	INRA	http://urgi.versailles.inra.fr/Wheat3BMine/	Wheat 3B data warehouse
12	IPK	http://lailaps.ipk-gatersleben.de	LAILAPS search engine
13	IPK	http://webblast.ipk-gatersleben.de/barley	IPK barley BLAST server
14	IPK	http://barlex.barleysequence.org	Barlex home page
15	TGAC	www.tgac.ac.uk/tools-resources	TGAC tools and resources homepage
16	TGAC	http://polymarker.tgac.ac.uk	TGAC Polymarker
17	TGAC	http://www.tgac.ac.uk/grassroots-genomics	TGAC grassroots genomics website
18	transPLANT	http://www.transplantdb.eu/training-resources	transPLANT user training resources
19	transPLANT	http://www.transplantdb.eu/videos	transPLANT user training videos
20	WheatIS	http://www.wheatinitiative.org/tools/wheat	Wheat Initiative website
21	WheatIS	www.wheatis.org	Wheat Information System homepage
22	PGSB	http://pgsb.helmholtz-muenchen.de/plant/plantsdb.jsp	PlantsDB entry page
23	PGSB	http://pgsb.helmholtz-muenchen.de/plant/barley/gz/tablejsp/index.jsp	PlantsDB GenomeZipper view
24	TGAC	http://browser.tgac.ac.uk/wheat/	TGAC wheat browser
25	TGAC	http://browser.tgac.ac.uk/barley_phys/	TGAC physical map browser
26	TGAC	http://browser.tgac.ac.uk/wheat_compara/	TGAC Aequatus browser

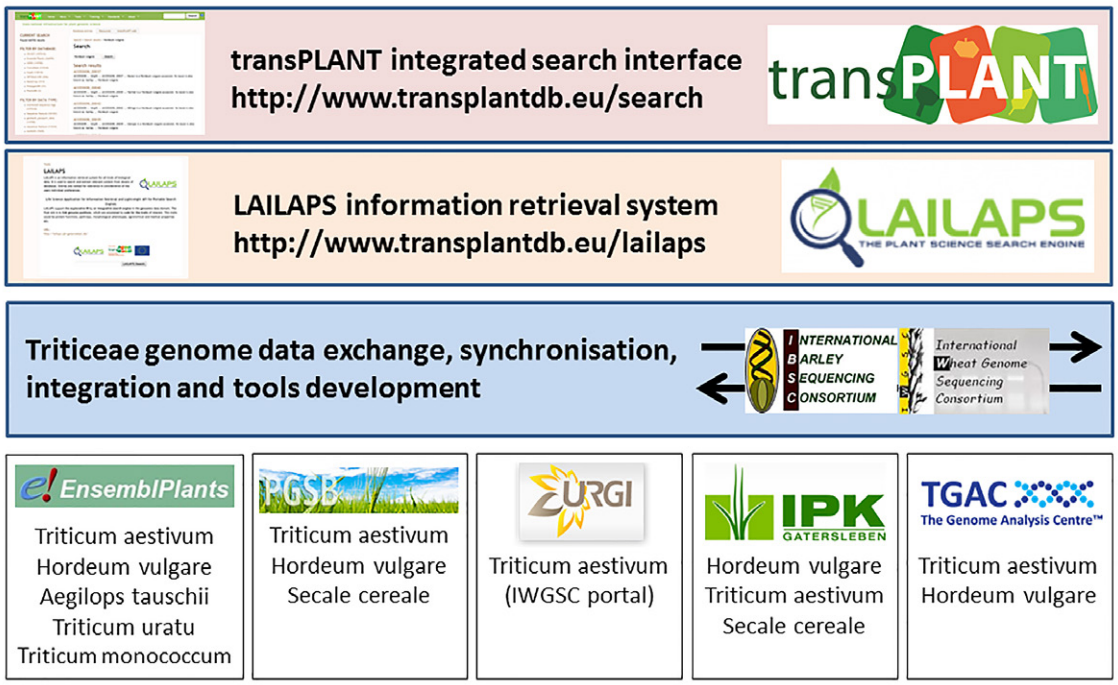

Figure 1. Overview of Triticeae genome resources and services provided by transPLANT partners. Species for which data is available in the respective database systems are given under the resource names. Details on data types and tools and modes of access are given in the individual partner sections. Upper panels illustrate both the transPLANT integrated search tool (enabling centralized keyword searches over the data inventories of many transPLANT partners) as well as the LAILAPS search engine (linking plant genomic data to phenotypic traits).

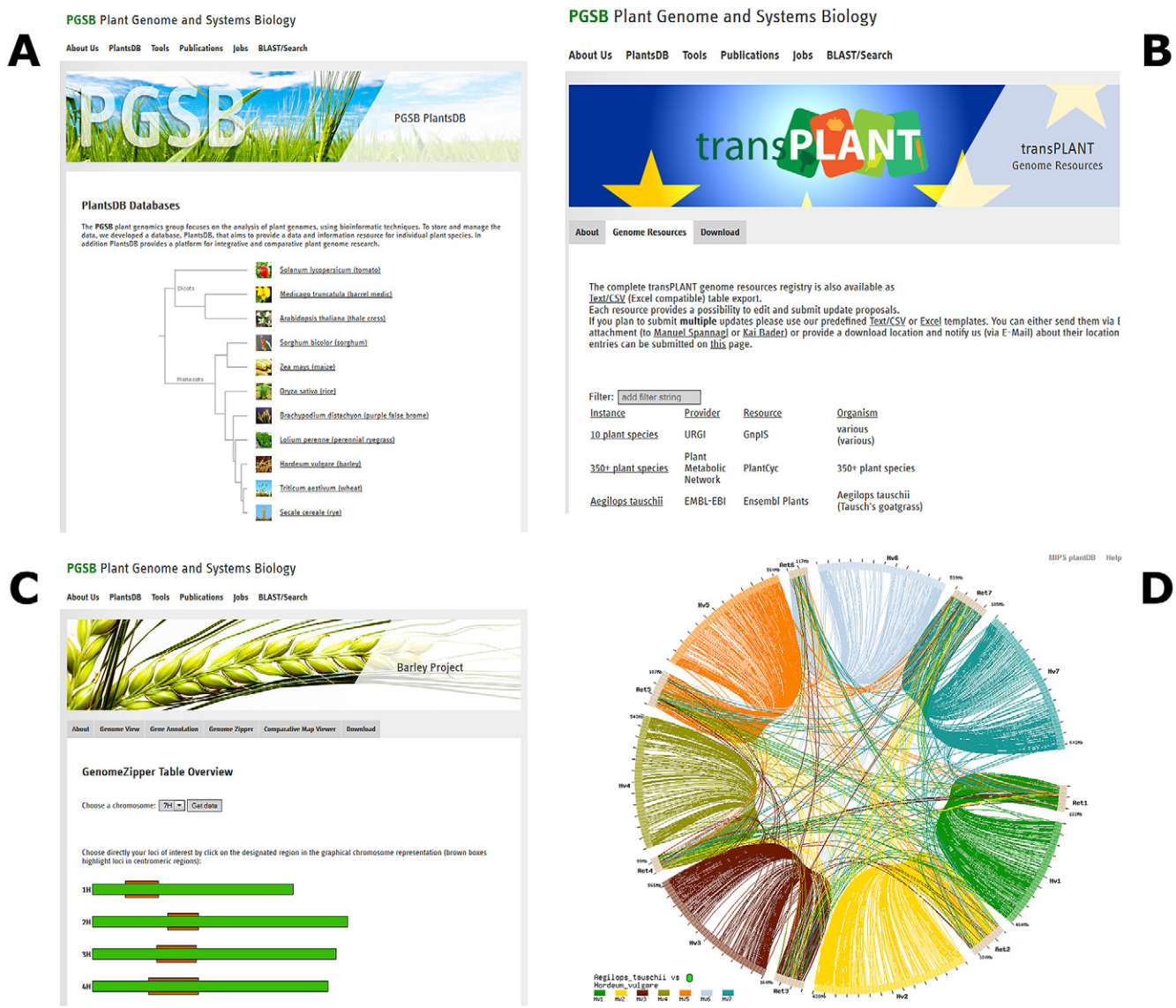

Figure 2. Triticeae and transPLANT resources hosted by PGSB PlantsDB. (A) Database overview and PlantsDB entry page (Table 1.22). (B) transPLANT genome resources registry giving access to more than 300 different plant genome resources and databases (Table 1.3). (C) GenomeZipper overview representation for the barley genome (Table 1.23). (D) Visualization of syntenic relationships between the genome sequences of *Aegilops tauschii* and barley within the CrowsNest tool (Table 1.5).

partner resources and the provision of web services and data warehouses. To assist users and communities in identifying data critical for their research, transPLANT provides a registry for genome resources (not restricted to Triticeae) and all data is discoverable within a single, integrated search. The registry can be accessed at the websites indicated in Table 1.2 and 1.3 (see also Fig. 2B). The integrated search is available from the website indicated in Table 1.1.

Plant Genome and Systems Biology PlantsDB: A Platform for the Comparative Analysis of Triticeae Genome Data

Plant Genome and Systems Biology (PGSB, formerly MIPS) PlantsDB provides a platform for the integration, visualization, and comparative analysis of plant genome data. Currently, genome data from 12 different plant species are available in the public domain of PGSB PlantsDB (Fig. 2A). Access points include search and browse interfaces, BioMoby webservices (Wilkinson et al., 2005), FTP download server, as well as visualization tools. With ongoing involvement in Triticeae genome sequencing and annotation initiatives such as IBSC (IBSC, 2012) and IWGSC (IWGSC, 2014), a major focus of PlantsDB was put on the representation and analysis of complex Triticeae genome data.

To compensate for the lack of reference chromosome sequences, GenomeZippers were constructed for many Triticeae species including barley, wheat, and rye. The GenomeZipper concept integrates data from chromosome sorting, second generation sequencing, array hybridization, and systematic exploitation of conserved

synteny with model grasses to construct virtual ordered gene maps (Mayer et al., 2009, 2011). Besides batch download via FTP, all GenomeZipper data has been integrated into the PGSB PlantsDB database scheme. To assist the interactive exploration of GenomeZippers and the search for anchored elements such as expressed sequence tags (ESTs), genetic markers, and full-length complementary DNA, interfaces for querying and browsing the GenomeZippers for barley, wheat, and rye were constructed (accessible from Table 1.4; Fig. 2C).

Between the genomes of many monocotyledonous plants, including major Triticeae crops and model plants, gene order appears to be conserved over long chromosomal stretches (synteny). This facilitates the potential transfer of knowledge from model plants such as *Brachypodium distachyon* (L.) Beauv. or rice (*Oryza sativa* L.) to the more complex Triticeae crops such as barley and wheat. To assist interactive exploration and visualization of syntenic regions and genes between grass models and Triticeae crop species, the CrowsNest tool (Table 1.5) was developed and populated with the genomes of rice, sorghum [*Sorghum bicolor* (L.) Moench], *B. distachyon*, barley, and *A. tauschii*, the diploid progenitor of the D sub-genome of hexaploid bread wheat. Figure 2D shows a screenshot from the CrowsNest tool visualizing synteny between *A. tauschii* and barley on a whole-genome scale.

Wheat genomic subassembly sequences generated in a reference-directed approach (Brenchley et al., 2012) were integrated with their corresponding genes from *B. distachyon*, sorghum, rice, and barley. Interfaces to query this data include a BLAST server to search for homologous wheat sequences as well as search for reference genes and associated wheat sequences. Genes predicted from chromosome-sorted wheat genome sequence generated within the IWGSC (IWGSC, 2014) have been integrated into PGSB PlantsDB and cross-referenced with the corresponding repositories GnpIS Wheat and Ensembl Plants.

To visualize and search the integrated barley physical and genetic maps, dedicated instances of GBrowse and CrowsNest were set up and in cooperation with IPK Gatersleben populated with markers, bacterial artificial chromosome (BAC) end sequences, BAC sequences, and physical map information. Gene expression data from barley (IBSC, 2012) has been integrated into the RNASeq-ExpressionBrowser (Nussbaumer et al., 2014) and can be queried by keyword, sequence similarity search (BLAST), or gene ontology (GO) and Interpro term and domain.

Triticeae Genome Data in Ensembl Plants

Ensembl Plants (Table 1.6) is an integrative web portal for plant genomic data, developed by the EBI. The portal provides interactive and programmatic access to data from 39 species through a variety of interfaces including web browser, Perl and RESTful Application Programming Interfaces, FTP, a publicly accessible relational database server, and a data-mining tool implemented using the data-warehousing framework BioMart optimized for gene and variant-centric queries. In addition to participating in transPLANT, coordination with efforts in the United States is achieved through a formal collaboration with the Gramene project (http://www.gramene.org).

Currently, four Triticeae genomes (among 20 cereal genomes) are available in Ensembl Plants: bread wheat, two of its diploid progenitors, *A. tauschii* and *T. urartu*, and barley. In the case of both wheat and barley; additional information (from genetic and physical maps) has been used to assign the genomic contigs to chromosomes and locate them within them. For barley, many of the contiguous sequences generated and assembled by the IBSC have been binned into located clusters according to evidence from the genetic and physical maps, and this information is used to construct a chromosome level view in Ensembl. The initial assembly has recently been revised using POPSEQ (Mascher et al., 2013a) data generated by the IPK. Whole-genome alignments have been performed against *B. distachyon*, rice, bread wheat, and the bread wheat progenitor genomes; collections of barley and wheat ESTs and RNA-sequencing (RNA-seq) reads have been aligned to the barley reference. In addition, intervarietal single nucleotide polymorphisms (SNPs) are represented for eleven varieties of barley as well as sites of variation between wild barley (*H. spontaneum*) and the barley reference.

The core wheat data represented is the chromosome survey sequence (CSS) generated, assembled, and annotated by the IWGSC. However, the CSS assembly of chromosome 3B has been replaced by the BAC-by-BAC assembly constructed by Choulet et al. (2014). In addition to sequence assemblies and gene models, a number of additional data sets have been aligned to the survey sequence, including the complete genomes of *B. distachyon* and rice, wheat unigene clusters from NCBI, and wheat RNA-seq data deposited in the International Nucleotide Sequence Database Collaboration archives. The wheat genome assemblies previously generated by Brenchley et al. (2012) have also been aligned to the survey sequence *B. distachyon* and barley. In addition, a collection of 900,000 polymorphisms from CerealsDB (http://www.cerealsdb.uk.net) have been included in the resource as well as data from the wheat HapMap project (Jordan et al., 2015). The chromosome survey sequence (and its annotations), in addition to the complete EST set, are available to search via BLAST and other search alignment algorithms.

For all gene models in Ensembl Plants, functional annotation is inferred using GO, InterPro, and homology metrics. Additionally, the evolutionary history of each protein-coding gene is inferred from comparisons to other plant species, and gene trees and protein alignments are available to browse (see Fig. 3) and download. The three bread wheat genomes have additionally been aligned to each other and linked to assertions of homeology derived from the gene tree analysis to provide supporting evidence. These alignments have been used to identify sites of variation (single nucleotide variants and insertion–deletions) between the A, B, and D genomes (see Fig. 4).

In addition, transcriptome data from another bread wheat precursor species, *T. monoccocum* (Fox et al., 2014), have been aligned to the hexaploid reference.

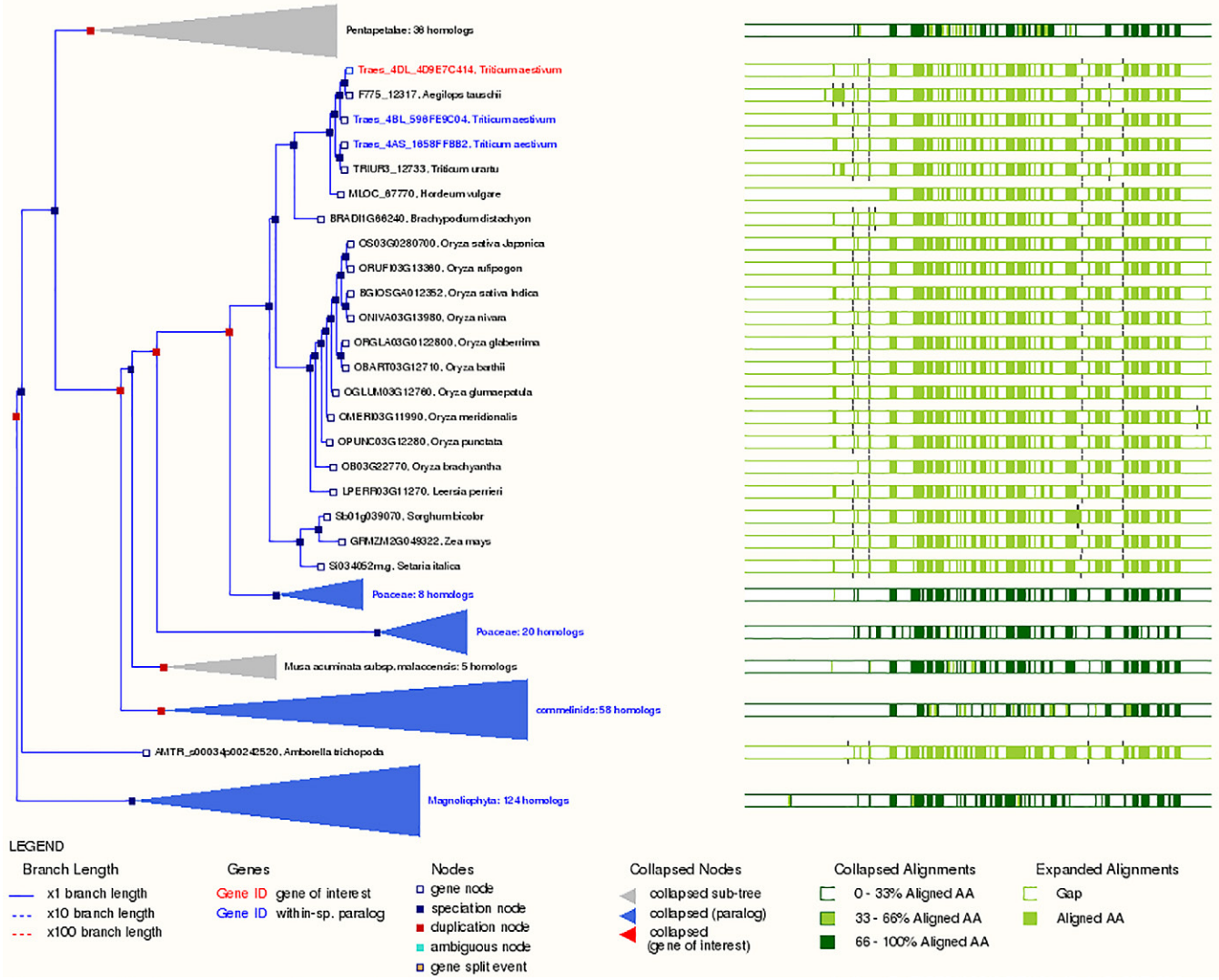

Figure 3. A gene family conserved with 1:1 orthology over 21 *Poaceae* genomes as visualized in Ensembl Plants. The three bread wheat genes are highlighted in red and blue, and the chromosome (and subgenome) are indicated in the prefix of the gene name (e.g., Traes_4AL_X is the name of a gene from the long arm of chromosome 4 in the A genome). The most related genes are those of the bread wheat precursors, followed by barley. Genes from more distant *Poaceae* species are located below.

Triticeae Genome Data and Tools at the Unité de Recherches en Génomique Info

The Official Wheat International Wheat Genome Sequencing Consortium Portal

Unité de Recherches en Génomique Info has been chosen by the IWGSC to be the repository for wheat genomic sequences and physical maps (Table 1.7).

To allow users to download, display, and query the IWGSC sequences and physical map data, a section dedicated to wheat genomics data, the sequence repository (Table 1.8; Fig. 5A), has been set up. Data stored in the sequence repository includes the wheat survey sequence, the chromosome reference sequence (chromosome 3B), the genes and annotations (gene models, GenomeZipper, and POPSEQ), the physical maps, the RNA-Seq, and the variations (HapMap) data.

Users can display the sequence annotation of the 3B reference sequence and the survey sequence in dedicated browsers. The physical maps browser (Table 1.9; Fig. 5B) is a customized instance of the GBrowse tool developed by the GMOD community (Stein et al., 2002).

The *T. aestivum* sequence data, diploid, and tetraploid wheat species sequence data (e.g., *T. durum*, *T. monococcum*, *T. urartu*, *A. speltoides*, *A. sharonensis*, and *A. tauschii*) are searchable using a BLAST tool (Table 1.10). The BLAST server is a customized version of the ViroBLAST tool developed by the University of Washington (Deng et al., 2007).

It also hosts the supplementary data attached to IWGSC publications and we are currently developing a page dedicated to assist the upcoming reference chromosome assemblies.

Data Warehouse for Wheat Chromosome 3B

To be able to connect reference sequence data from chromosome 3B (Choulet et al., 2014) with genetics and phenomics data, the Wheat3BMine data warehouse (Table 1.11; Fig. 5C) was developed in the framework of transPLANT.

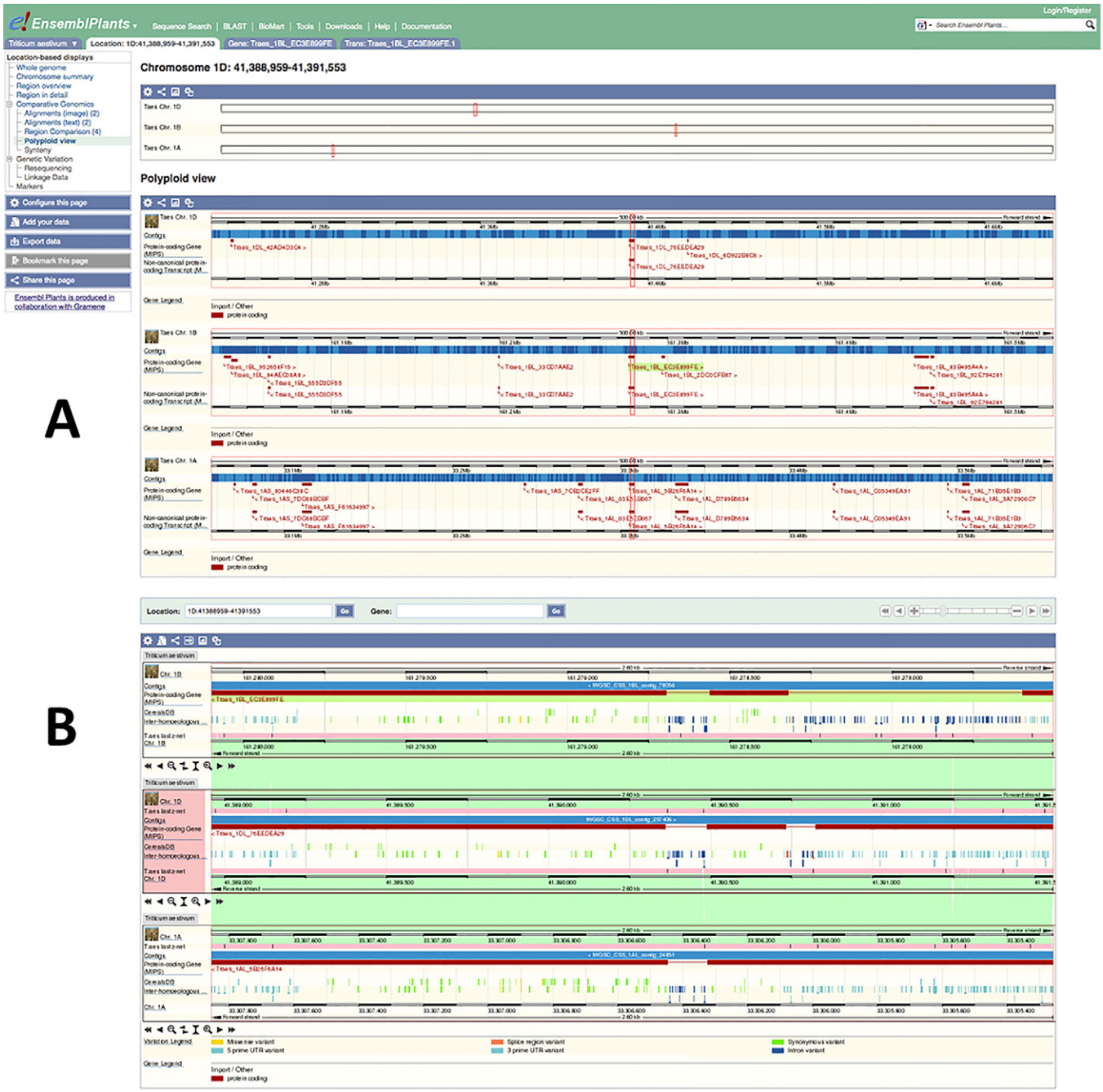

Figure 4. Homeologous regions from the bread wheat A, B, and D genomes as visualized in Ensembl Plants. (A) The top panel shows the annotations made on three regions of contiguous sequence. (B) The lower panel is centered on the homeologous genes and shows gene structures, intervarietal polymorphisms and interhomeologous variants.

The warehouse is implemented using InterMine technology that provides a fast, flexible, and user friendly access to integrated data by multiple ways: a browser, a query builder, and a region search tool. Wheat3BMine users can filter their favorite features, save their own queries, and export results in many different formats (GFF3, BED, or XML). An online documentation and precomputed queries are also available.

The data warehouse contains heterogeneous data and is gene centric. In fact, the typical gene card centralizes relevant information like gene function, ontology terms, and overlapping features. Wheat3BMine provides access to genomic annotation data (genes, mRNA, polypeptides, and transposable elements), polymorphisms data (markers), genetic mapping data (quantitative trait loci [QTL], meta-QTL), and phenotyping data. Moreover, useful links are available from a gene card to the wheat 3B genome browser (Choulet et al., 2014) and to additional details in GnpIS.

Wheat Data in the GnpIS Information System

GnpIS (Steinbach et al., 2013) is an information system that integrates genomic and genetic data for plants and fungi. A "wheat" filter was implemented within GnpIS that allows interconnecting the wheat genomic data

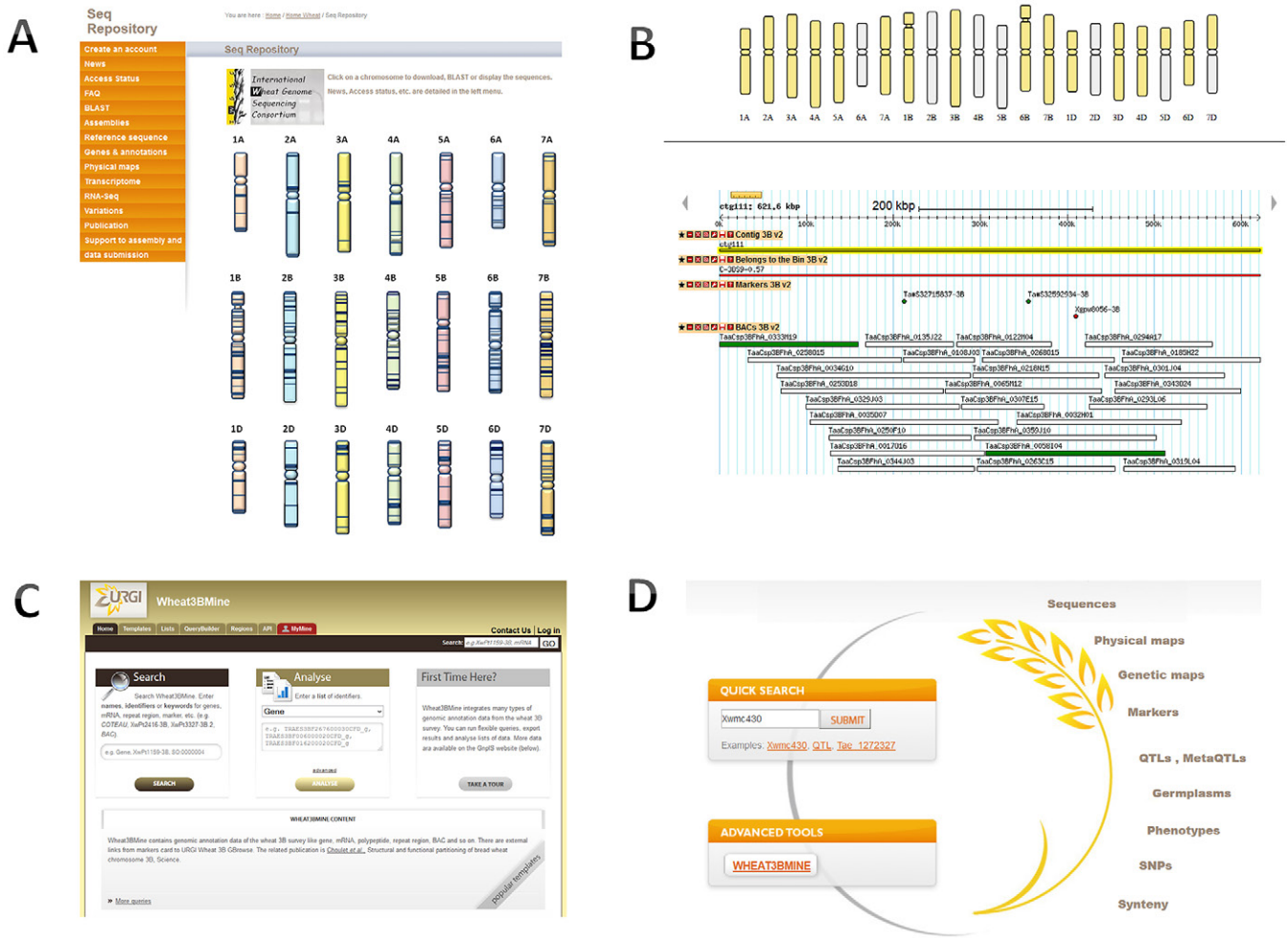

Figure 5. Triticeae resources hosted by URGI. (A) Wheat@URGI website: IWGSC Sequence Repository page (Table 1.8). (B) Wheat physical maps browser (Table 1.9). (C) Wheat3BMine tool homepage (Table 1.11). (D) Wheat data search in GnpIS: quick search, advanced tool and dedicated web interfaces (Table 1.7).

detailed above with the germplasm, markers, QTLs, SNPs, expression, and phenotypes data. Moreover, association and genomic selection data are in the process of integration into the information system. The wheat data in GnpIS (Table 1.7; Fig. 5D) can be queried using the quick search tool (Google-like search), advanced search tool (Wheat3BMine), and the dedicated web interfaces developed in Java and Google Web Toolkit.

Triticeae Genome Data and Tools at the Leibniz Institute of Plant Genetics and Crop Plant Research Gatersleben

LAILAPS Search Engine

LAILAPS (Esch et al., 2014) is an integrated information retrieval system to link plant genomic data in the context of phenotypic traits for a detailed forward genetic research (Table 1.12; Fig. 6). LAILAPS is developed in the framework of the transPLANT project and allows exploratory search for candidate genes linked to specific traits over a loosely integrated system of indexed and interlinked genome databases. Query assistance and an evidence-based annotation system enable time-efficient and comprehensive information retrieval. An artificial neural network incorporating user feedback and behavior tracking allows relevance sorting of results. Because this enhanced relevance ranking is one of the major innovations to explore millions of database records, a special focus has been set to its training and the inclusion of user feedback. The current LAILAPS release comprises about 91 million indexed database records of trait knowledge within 13 major life science data collections and more than 60 million associations to -omics data sets. To provide an up-to-date user ergonometry, the front end features an interactive query assistance that suggests spelling correction as well as semantic query expansion. This feature makes use of PubMed abstracts to learn vectors of similar words and phrases. Queries are expressed as keyword or phrases that are spell corrected. As query results, a condensed list of relevant hits is rendered that includes a short excerpt of relevant text positions and a list of annotation links to annotated genes in plant genomes. A comprehensive result filter panel and the suggestion of semantic follow-up queries rounds off the

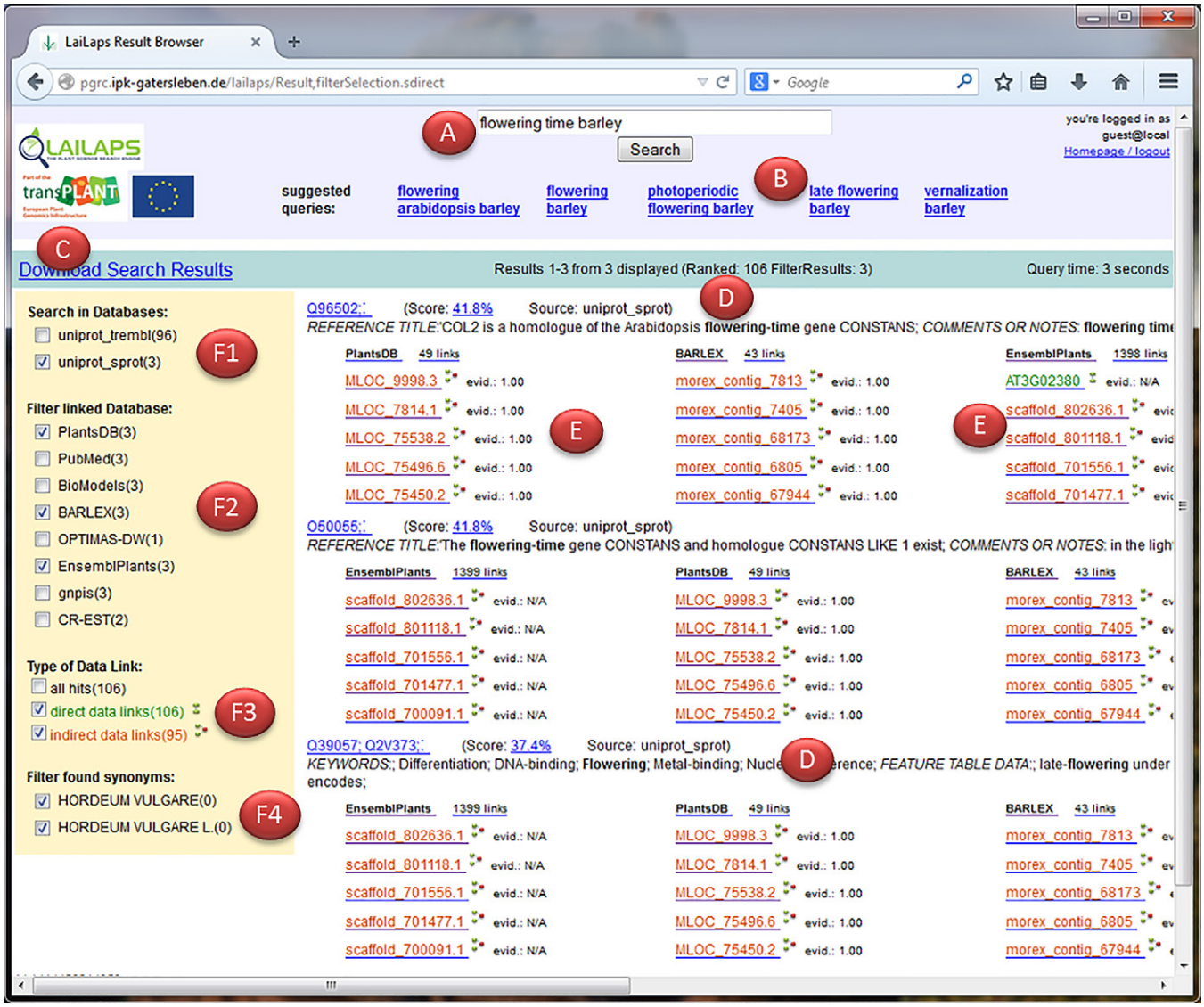

Figure 6. The LAILAPS web interface, illustrating the search and results page. The text box (A) enables an interactive and spelling corrected keyword query submission. After query was executed successfully, a list of semantically related phrases is provided for query refinement (B). The query results can be either downloaded as a Microsoft Excel sheet (C) or interactively explored. For this, all relevant hits are displayed as short excerpts in the result panel (D). Connected to each hit is a list of links to associated genomic data (E). Those links can either refer to genome data directly (green) or reflect an indirect, transitive relationship (red). The left hand filter panel enables to restrict the results by fact databases (F1), linked genome databases (F2), direct or indirect linked gene annotations (F3), or synonyms (F4).

query refinement features. Once the user identifies interesting records, he is guided to the most relevant genomic dataset and is able to rate its quality. In turn, this enables LAILAPS relevance prediction system to improve the relevance prediction network.

Barley BLAST Server

The IPK hosts a BLAST server (Table 1.13), providing homology searches against the complete barley sequence data sets published in IBSC (2012). This comprises the whole-genome shotgun assemblies of the cultivars Morex, Barke, and Bowman as well as the high- and low-confidence gene sets. The latest POPSEQ anchoring data (Mascher et al., 2013a) was also integrated in addition to the barley exome capture targets (Mascher et al., 2013b).

BARLEX

The recent progress in sequencing and mapping technology has facilitated the construction of advanced genomics resources in species with large and complex crop genomes like barley. During such genome sequencing projects, the integration of large volumes of diverse information and data from disparate sources is an open issue. Existing genome browsers are not well adapted to this task, as they expect all genomic features to be anchored to a single linearly ordered reference sequence. The IPK provides the barley genome explorer, BARLEX (Colmsee et al., 2015), as a central unified repository for the genomic resources of barley. BARLEX is centered on the genome-wide physical map of barley and links it to an annotated whole-genome shotgun assembly and dense genetic

maps. A web-based interface presents data in tabular and graphical format and associates all information and published sequence data with shotgun assemblies, repeat annotations using KMasker (see Schmutzer et al., 2014), physical contigs, and annotated genes. A novel graph-based visualization strategy was implemented to show overlaps between adjacent BACs based on fingerprint and sequence data. BARLEX is publicly accessible at the website listed in Table 1.14 and is directly connected to the IPK Barley BLAST server as well as the LAILAPS system.

e!DAL: Plant Genomics and Phenomics Research Data Repository

The IPK is hosting a plant genomics and phenomics research data repository, which is based on the e!DAL data sharing and publication infrastructure (Arend et al., 2014). It features the publication of plant research data that is out of scope of existing domain databases, too huge, or less structured. In compliance to international standards, such as DOI, DataCite, and OpenAIRE, plant genomic and phenotpyic datasets are published. In a particular focus are studies of plant genetic resources from the system plant from the root to bloom and seed, as well from sequence analysis to systems biology. Examples are genomic datasets of Triticeae (DOIs: 10.5447/IPK/2015/0, 10.5447/IPK/2015/1, and 10.5447/IPK/2015/2).

Triticeae Genome Data and Tools at The Genome Analysis Centre

The TGAC Browser is an open-source genomic browser developed to visualize genome annotations such as genes, variations, and markers for species whose reference sequence may be contiguous or highly fragmented; the IBSC's barley genome and the wheat CSS represent two such fragmented references (Fig. 7A, 7B). Traditional datatypes, such as the reference assembly and gene annotations, reside in an Ensembl schema database. Larger datasets are stored in standard file formats such as SAM, BAM, bigwig, and VCF. The TGAC Browser has an integrated BLAST functionality, essential for identifying and accessing regions of interest in fragmented genomes, and an interface to enable manual annotation of genomic features. Homeologous genes can also be explored through the Aequatus browser (Fig. 7C).

Through collaboration with CerealsDB, the TGAC wheat browser displays the mapped SNP markers from the 90K iSelect and Axiom arrays against the IWGSC chromosome survey sequence contigs. The TGAC Browsers are also available for the barley genome and barley physical map, which display the minimum tile path and BAC-end sequences. All browsers can be accessed at Table 1.15.

To tackle the issue of marker design for polyploid genomes, we have developed PolyMarker, an automated bioinformatics pipeline for SNP assay development that increases the probability of generating homeologue-specific assays for polyploid wheat (Ramirez-Gonzalez et al., 2015). PolyMarker (Fig. 7D) generates a multiple alignment between the target SNP sequence and the IWGSC chromosome survey sequences (IWGSC, 2014) for each of the three wheat genomes. It then generates a mask with informative positions, which are highlighted with respect to the target genome allowing homeologue-specific primer design. The PolyMarker site (Table 1.16) provides predesigned primers for the iSelect 90K chip and 820K Axiom markers.

For more information and community support for these resources TGAC hosts the Grassroots genomics website (Table 1.17).

Discussion

Outlook

Great progress has been made recently in sequencing, annotating, and analyzing the complex genomes of Triticeae species including bread wheat and barley. The data generated has great potential for applications in plant breeding, experimental plant biology, and comparative genomics. However, to make the best possible use of the large and heterogeneous data sets produced, data archiving, integration, visualization, and access are essential. A variety of platforms provide different types of analysis and presentations, but it can be difficult for users to use these resources in combination. Many of the partners within the transPLANT project are actively involved in past or ongoing Triticeae genome initiatives and the development of resources. A major focus of the project is ensuring interoperability to maximize their collective value.

Critical to this is the development of common standards and formats. transPLANT partners use accepted standards to share and disseminate data wherever possible and are involved in ongoing efforts to standardize plant phenotypic data and metadata (that is, the description of material, experimental conditions, and results [Krajewski et al., 2015]). We have also been working with common data mining interfaces (BioMart [Smedley et al., 2015], InterMine [Smith et al., 2012]), and developing RESTful web services to support integrative programmatic access to data. We have also been developing cloud computing environments to support downstream analyses. This has translated into a number of concrete benefits for end users working with complex Triticeae genome data, including the following:

- Extensive and standardized data exchange and synchronization between partners; all data is served on common reference sequence, and portable annotation tracks can be visualized at different sites
- Data retrieval tool at the transPLANT web hub, indexing multiple types of Triticeae genome data (e.g., ESTs, genes, transcripts, phenotypes, and accessions) from all transPLANT partners
- LAILAPS integrated search engine, linking various Triticeae genomic data from transPLANT partners in the context of phenotypic traits

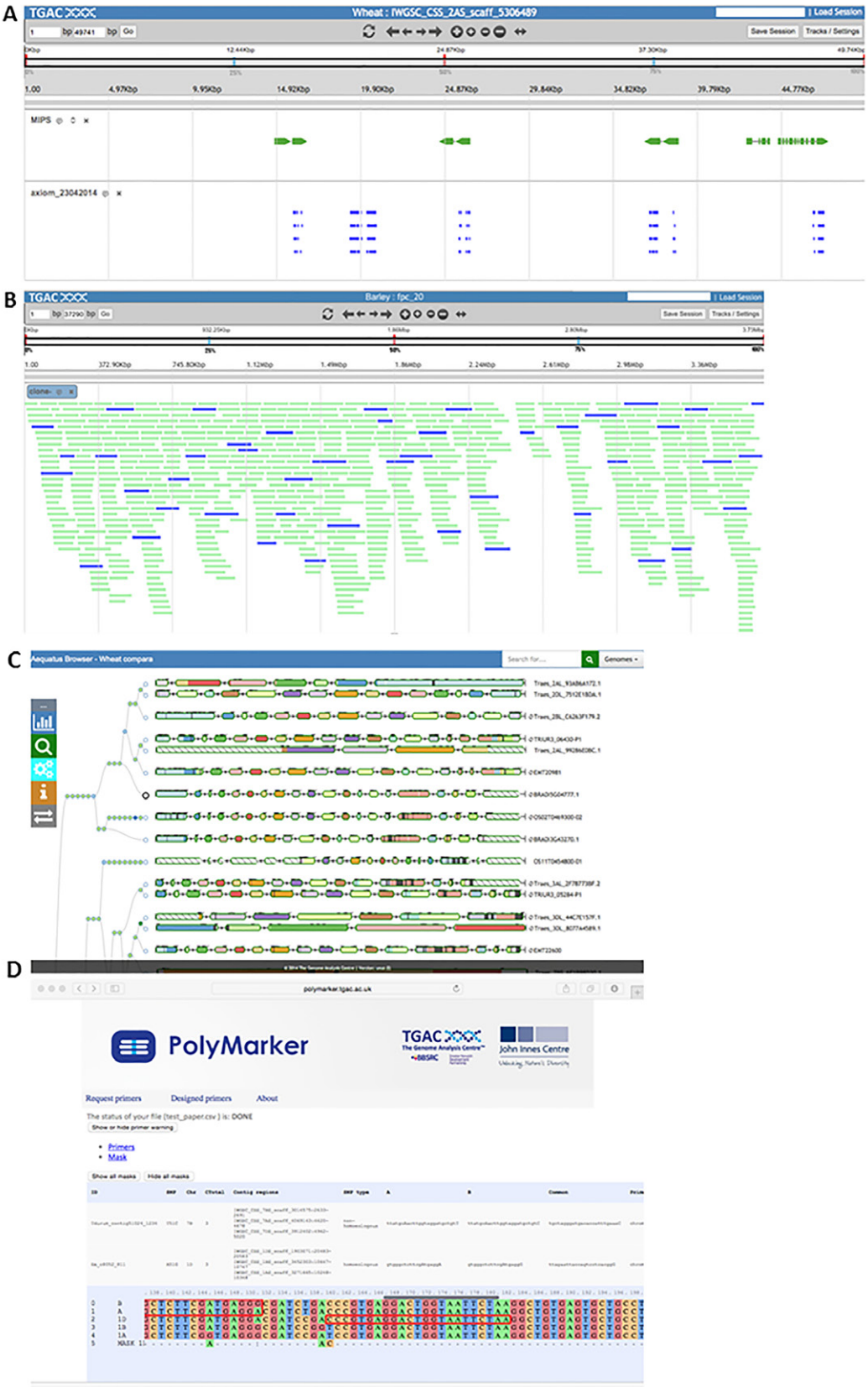

Figure 7. The Genome Analysis Centre (TGAC) wheat genome resources. (A) Wheat TGAC Browser: TGAC browser (Table 1.24) showing *Triticum aestivum* scaffold IWGSC_2AS_scaff_5306489, with SNPs from 820K Axiom array and gene annotations from MIPS/PGSB. (B) TGAC Browser Barley Physical Map: TGAC browser showing the barley physical map for fingerprint (FPC) contig 20 (Table 1.25). MTP BACs are highlighted in blue. (C) Aequatus Browser (Table 1.26) showing *Brachypodium distachyon* gene BRADI5G04777.1 and homologous genes from *Triticum aestivum*, *Aegilops tauschii*, *Oryza sativa*, and *Triticum uratu*. (D) Primers designed by PolyMarker (Table 1.16). The webpage highlights the designed primers for validation that can be downloaded in spreadsheet formats.

- Extensive cross-linking between Triticeae genome resources and tools
- Provision of programmatic access to databases via APIs or web services
- Comprehensive online user training material available for download, covering both resource usage and wider analytical approaches, developed during a series of hands-on workshops (Table 1.18)
- Integrated training videos providing an overview of available resources (Table 1.19)

The result of these efforts is a rich collection of interfaces using common reference data, searchable through a single entry point at the central transPLANT web hub. They also simplify the identification of suitable datasets and databases for research using Triticeae genome data, assist in data acquisition, and provide powerful tools to analyze data in the context of other plant species. The training videos provide example use cases illustrating how users can take advantage of different resources in combination to interrogate the data and perform complex analyses.

With the expected emergence of additional genome sequence data from the Triticeae, the framework for data integration, exchange, and aggregation established within the transPLANT project will help to address the challenges involved with even more distributed data repositories and heterogeneous data types. In this context, a project has started recently with the objective to set up an International Wheat Information System (WheatIS) to support the wheat research community. The main objective is to provide a single-entry web-based system to access the available data resources and bioinformatics tools. The WheatIS project is an international project lead by an Export Working Group of the Wheat Initiative (Table 1.20). The Wheat Initiative is supported by the G20 Agricultural Ministers to coordinate worldwide research efforts in the fields of wheat genetics, genomics, physiology, breeding and agronomy.

The WheatIS project is driven by a network of 21 experts from Australia, Canada, France, Germany, Mexico, United States, and United Kingdom that congregates a group of volunteers willing to participate in the WheatIS project. The WheatIS (Table 1.21) will operate as a hub integrating wheat data produced and submitted to the public repositories by the community, extending the model and technologies established in transPLANT for the coordination of dispersed resources.

Acknowledgments

All authors and institutions would like to acknowledge funding of the transPLANT project by the European Commission within its 7th Framework Programme, under the thematic area Infrastructures, contract number 283496. EBI acknowledges funding from the United Kingdom Biotechnology and Biological Sciences Research Council grants BB/I008071/1 and BB/J00328X/1, and grant 52930112 from the United States National Science Foundation. The IPK resources Barley Blast Server, BARLEX, and Kmasker e!DAL were supported by the Leibniz Association (WGL) in the context of the *Pakt für Forschung und Innovation*/WGL and the German Federal Ministry of Education and Research (BMBF) in the frame of the projects BARLEX (FKZ 0314000A), and TRITEX (FKZ 0315954A) and DPPN (FKZ 031A053B). URGI likes to acknowledge funding from INRA, french Research National Agency (ANR-09-GENM-025-003) 3BSEQ project, Investment for the Future (ANR-10-BTBR-03, France AgriMer, FSOV) BreedWheat project, European commission within 7th Framework Program TriticeaeGenome (KBBE-212019) and WHEALBI (FP7-613556) projects. TGAC likes to acknowledge funding from the Biotechnology and Biological Sciences Research Council grants BB/L002124/1 and BB/L024144/1. RRG is supported by a Norwich Research Park PhD Studentship and The Genome Analysis Centre Funding and Maintenance Grant. PGSB acknowledges funding from the German Federal Ministry of Education and Research (BMBF) in the frame of the projects BARLEX (FKZ 0314000A), and TRITEX (FKZ 0315954A) as well as Deutsche Forschungsgemeinschaft (DFG) funding to project SFB924 Molecular mechanisms regulating yield and yield stability in plants.

References

Arend, D., M. Lange, J. Chen, C. Colmsee, S. Flemming, D. Hecht, et al. 2014. e!DAL: A framework to store, share and publish research data. BMC bioinformatics 15: 214. doi:10.1186/1471-2105-15-214

Brenchley, R., M. Spannagl, M. Pfeifer, G.L. Barker, R. D'Amore, A.M. Allen, et al. 2012. Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 491:705–710. doi:10.1038/nature11650

Choulet, F., A. Alberti, S. Theil, N. Glover, V. Barbe, J. Daron, et al. 2014. Structural and functional partitioning of bread wheat chromosome 3B. Science 345:1249721. doi:10.1126/science.1249721

Colmsee, C., S. Beier, A. Himmelbach, T. Schmutzer, N. Stein, U. Scholz, et al. 2015. BARLEX: The Barley Draft Genome Explorer. Mol. Plant 8:964–966. doi:10.1016/j.molp.2015.03.009

Deng, W., D.C. Nickle, G.H. Learn, B. Maust, and J.I. Mullins. 2007. ViroBLAST: A stand-alone BLAST web server for flexible queries of multiple databases and user's datasets. Bioinformatics 23:2334–2336. doi:10.1093/bioinformatics/btm331

Esch, M., J. Chen, C. Colmsee, M. Klapperstuck, E. Grafahrend-Belau, U. Scholz, et al. 2014. LAILAPS: The plant science search engine. Plant Cell Physiol. 56:e8. doi:10.1093/pcp/pcu185

Fox, S.E., M. Geniza, M. Hanumappa, S. Naithani, C. Sullivan, J. Preece, et al. 2014. De novo transcriptome assembly and analyses of gene expression during photomorphogenesis in diploid wheat *Triticum monococcum*. PLoS ONE 9:E96855. doi:10.1371/journal.pone.0096855

International Barley Genome Sequencing Consortium. 2012. A physical, genetic and functional sequence assembly of the barley genome. Nature 491:711–716. doi:10.1038/nature11543

International Wheat Genome Sequencing Consortium. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. Science 345:1251788. doi:10.1126/science.1251788

Jia, J., S. Zhao, X. Kong, Y. Li, G. Zhao, W. He, et al. 2013. *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. Nature 496:91–95. doi:10.1038/nature12028

Jordan, K.W., S. Wang, Y. Lun, L.J. Gardiner, R. MacLachlan, P. Hucl, et al. 2015. A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. Genome Biol 16:48. doi:10.1186/s13059-015-0606-4

Krajewski, P., D. Chen, H. Cwiek, A.D. van Dijk, F. Fiorani, P. Kersey, et al. 2015. Towards recommendations for metadata and data handling in plant phenotyping. J. Exp. Bot. 66:5417–5427. doi:10.1093/jxb/erv271

Ling, H.Q., S. Zhao, D. Liu, J. Wang, H. Sun, C. Zhang, et al. 2013. Draft genome of the wheat A-genome progenitor *Triticum urartu*. Nature 496:87–90. doi:10.1038/nature11997

Luo, M.C., Y.Q. Gu, F.M. You, K.R. Deal, Y. Ma, Y. Hu, et al. 2013. A 4-gigabase physical map unlocks the structure and evolution of the complex genome of Aegilops tauschii, the wheat D-genome progenitor. Proc. Natl. Acad. Sci. USA 110:7940–7945. doi:10.1073/pnas.1219082110

Martis, M.M., R. Zhou, G. Haseneyer, T. Schmutzer, J. Vrana, M. Kubalakova, et al. 2013. Reticulate evolution of the rye genome. Plant Cell 25:3685–3698. doi:10.1105/tpc.113.114553

Mascher, M., G.J. Muehlbauer, D.S. Rokhsar, J. Chapman, J. Schmutz, K. Barry, et al. 2013a. Anchoring and ordering NGS contig assemblies

by population sequencing (POPSEQ). Plant J. 76:718–727. doi:10.1111/tpj.12319

Mascher, M., T.A. Richmond, D.J. Gerhardt, A. Himmelbach, L. Clissold, D. Sampath, et al. 2013b. Barley whole exome capture: A tool for genomic research in the genus *Hordeum* and beyond. Plant J. 76:494–505 10.1111/tpj.12294. doi:10.1111/tpj.12294

Mayer, K.F., M. Martis, P.E. Hedley, H. Simkova, H. Liu, J.A. Morris, et al. 2011. Unlocking the barley genome by chromosomal and comparative genomics. Plant Cell 23:1249–1263. doi:10.1105/tpc.110.082537

Mayer, K.F., S. Taudien, M. Martis, H. Simkova, P. Suchankova, H. Gundlach, et al. 2009. Gene content and virtual gene order of barley chromosome 1H. Plant Physiol. 151:496–505. doi:10.1104/pp.109.142612

Nussbaumer, T., K.G. Kugler, K.C. Bader, S. Sharma, M. Seidel, and K.F. Mayer. 2014. RNASeqExpressionBrowser: A web interface to browse and visualize high-throughput expression data. Bioinformatics 30:2519–2520. doi:10.1093/bioinformatics/btu334

Ramirez-Gonzalez, R.H., C. Uauy, and M. Caccamo. 2015. PolyMarker: A fast polyploid primer design pipeline. Bioinformatics 31:2038–2039. doi:10.1093/bioinformatics/btv069

Schmutzer, T., L. Ma, N. Pousarebani, F. Bull, N. Stein, A. Houben, et al. 2014. Kmasker: A tool for in silico prediction of single-copy FISH probes for the large-genome species *Hordeum vulgare*. Cytogenet. Genome Res. 142:66–78. doi:10.1159/000356460

Smedley, D., S. Haider, S. Durinck, L. Pandini, P. Provero, J. Allen, et al. 2015. The BioMart community portal: An innovative alternative to large, centralized data repositories. Nucleic Acids Res. 43:W589–W598. doi:10.1093/nar/gkv350

Smith, R.N., J. Aleksic, D. Butano, A. Carr, S. Contrino, F. Hu, et al. 2012. InterMine: A flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics 28:3163–3165. doi:10.1093/bioinformatics/bts577

Stein, L.D., C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, et al. 2002. The generic genome browser: A building block for a model organism system database. Genome Res. 12:1599–1610. doi:10.1101/gr.403602

Steinbach, D., M. Alaux, J. Amselem, N. Choisne, S. Durand, R. Flores, et al. 2013. GnpIS: An information system to integrate genetic and genomic data from plants and fungi. Database 2013:Bat058 doi:10.1093/database/bat058

Wilkinson, M., H. Schoof, R. Ernst, and D. Haase. 2005. BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. The PlaNet exemplar case. Plant Physiol. 138:5–17. doi:10.1104/pp.104.059170