The Plant Orthology Browser: An Orthology and Gene-Order Visualizer for Plant Comparative Genomics

Dan Tulpan* and Serge Leger

Abstract

Worldwide genome sequencing efforts for plants with medium and large genomes require identification and visualization of orthologous genes, while their syntenic conservation becomes the pinnacle of any comparative and functional genomics study. Using gene models for 20 fully sequenced plant genomes, including model organisms and staple crops such as *Aegilops tauschii* Coss., *Arabidopsis thaliana* (L.) Heynh., *Brachypodium distachyon* (L.) Beauv., turnip (*Brassica rapa* L.), barley (*Hordeum vulgare* L.), rice (*Oryza sativa* L.), sorghum [*Sorghum bicolor* (L.) Moench], wheat (*Triticum aestivum* L.), red wild einkorn (*Triticum urartu* Tumanian ex Gandilyan), and maize (*Zea mays* L.), we computationally predicted 1,021,611 orthologs using stringent sequence similarity criteria. For each pair of plant species, we determined sets of conserved synteny blocks using strand orientation and physical mapping. Gene ontology (GO) annotations are added for each gene. Plant Orthology Browser (POB) includes three interconnected modules: (i) a gene-order visualization module implementing an interactive environment for exploration of gene order between any pair of chromosomes in two plant species, (ii) a synteny visualization module providing unique interactive dot plot representations of orthologous genes between a pair of chromosomes in two distinct plant species, and (iii) a search module that interconnects all modules via free-text search capability with online as-you-type suggestions and highlighting that allows exploration of the underlining information without constraint of interface-dependent search fields. The POB is a web-based orthology and annotation visualization tool, which currently supports 20 completely sequenced plant species with considerably large genomes and offers intuitive and highly interactive pairwise comparison and visualization of genomic traits via gene orthology.

Core Ideas

- Innovative web-based orthology and annotation visualization tool
- Intuitive, highly interactive pairwise comparison and visualization of genomic traits
- Interactive synteny plots for exploration and discovery of gene impact/localization
- Free-text, Google-like search with as-you-type suggestions and highlighting
- Easy retrieval of orthologs, comparative results, and publication-quality figures

ORTHOLOGS are homologous genes that have diverged after a speciation event and typically retain similar functions across species, while paralogs are homologous genes that have diverged after a duplication event and have different functions. As a result of recent technological advancements in developing high-throughput and highly accurate sequencing technologies (which, in turn, produced a rapid increase in genomic data generation) identifying and distinguishing orthologous genes among different species becomes a quintessential step for comparative genomics and genome annotation. This is particularly relevant in plant genomics because of the enormous genome sizes of key staple crop species and other species of economic and scientific interest (Fig. 1). For example, the hexaploid bread wheat genome, which consists of three subgenomes (A, B, and D), is estimated

Information and Communications Technologies, National Research Council Canada, Moncton, New Brunswick, E1A 7R1, Canada. Received 9 Aug. 2016. Accepted 1 Dec. 2016. *Corresponding author (dan.tulpan@nrc-cnrc.gc.ca).

Abbreviations: API, application program interface; GO, gene ontology; POB, Plant Orthology Browser; RBBH, reciprocal best BLAST hit; WYSIWYG, what you see is what you get.
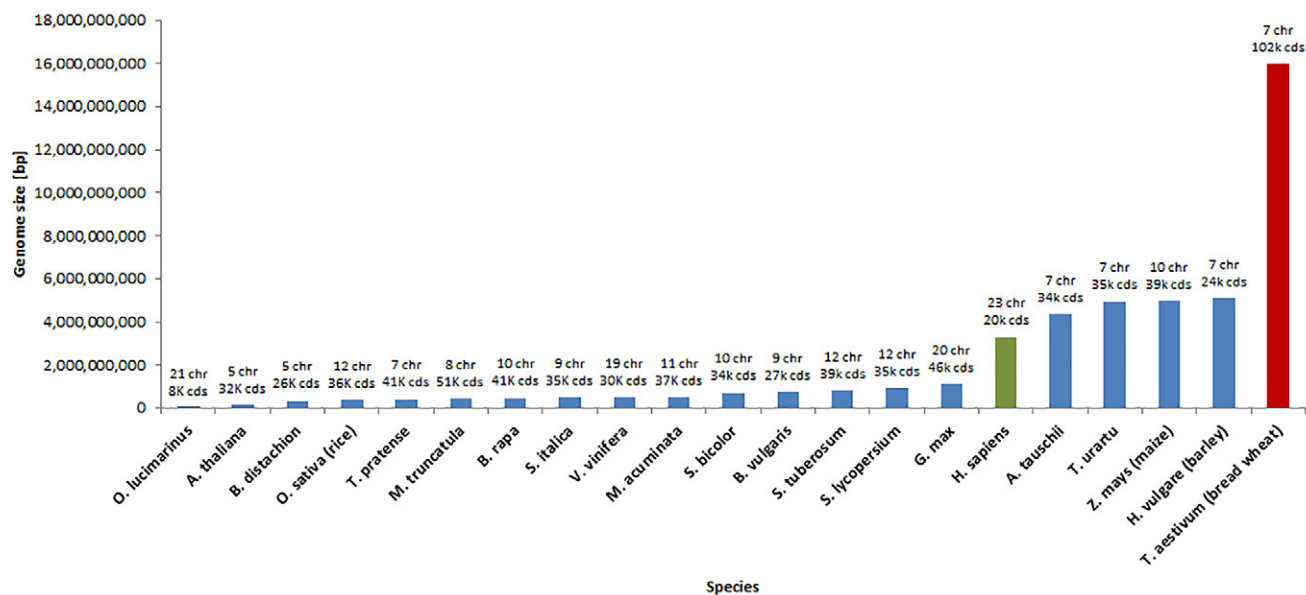
Estimated genome sizes

Fig. 1. Genome sizes of 20 plant species (blue bars, red bar for bread wheat) and the human genome (green bar). Data was obtained from Ensembl (*H. sapiens*) and Ensembl Plants (20 plant species).

to be approximately five times larger than the human genome (Kersey et al., 2015).

Orthology prediction methods can be generally grouped into two categories: (i) phylogenetic or tree-based methods that cluster genes and aim for the reconciliation of the protein and the species trees (e.g., TreeFam [Ruan et al., 2008], Ensembl Compara [Flicek et al., 2014], PhylomeDB [Huerta-Cepas et al., 2014], LOFT [van der Heijden et al., 2007], MetaPhOrs [Pryszcz, et al., 2011], and Otholuge [Fulton et al., 2006]) and (ii) pairwise sequence alignment or graph-based methods that cluster pairs of genes based on (typically protein) sequence similarity (e.g., InParanoid [Ostlund et al., 2010], RoundUp [Deluca et al., 2006], COG [Tatusov, et al., 1997], KOG [Tatusov et al., 2003], eggNOG [Powell et al., 2014], OrthoDB [Waterhouse et al., 2013], OrthoMCL [Chen et al., 2006], OMA [Altenhoff et al., 2011], PlantOrDB [Li et al., 2015], OrthoVenn [Wang et al., 2015], and SPOCS [Curtis et al., 2013]). While no group of orthology prediction methods holds a clear advantage over the other, and their usage is typically dependent on the requirements of the problem that must be addressed, systematic descriptions and comparisons of these methods including advantages and pitfalls can be found in (Altenhoff and Dessimoz, 2009; Fang et al., 2010). Phylogenetic methods typically consist of four steps: (i) homologous gene clustering, (ii) multiple sequence alignment, (iii) phylogenetic tree reconstruction, and (iv) detection of orthology and paralogy. While phylogenetic methods tend to be robust and accurate, they do not scale well (computationally) with the number of input sequences and are sensitive to the strategies applied in each of the four steps. To alleviate some of the challenges associated to phylogeny-based orthology prediction, particularly the computational scalability with number of input sequences, pairwise

sequence alignment methods were developed. Relying on the assumption that, in general, sequence similarity is a strong indication of orthology among genes or proteins, these methods are typically performed using a reciprocal best BLAST hit (RBBH) approach. While RBBH detection is computationally fast, it has its own drawbacks, such as being affected by incomplete genome sequences or gene loss, therefore incorrectly predicting a paralog as an ortholog. Nevertheless, in particular circumstances where trusted orthologs are required, such as information enrichment in genome annotation (Wolf and Koonin, 2012) and genome zipper construction and assessment (Silvar et al., 2015), the RBBH method generates high quality one-to-one orthologs, which can be further used to seed orthologous groups (Dalquen and Dessimoz, 2013).

Here, we present the POB (http://nrcmonsrv01.nrc.ca/pob/), an interactive web-based gene orthology and annotation visualization browser that uses 1,021,611 predicted orthologs from 20 plant species. The POB provides gene-order and synteny visualization capabilities of orthologous genes and their functional annotations. For any given gene, the user can infer its orthologs in different plant species through a highly responsive web interface that provides integrated, interactive, and interconnected visualization of gene-order and synteny blocks. In particular, POB allows users to explore all orthologs between pairs of chromosomes in two plants, which are predicted using an RBBH approach. The POB predicts orthologs based on two complementary methods: RBBHs using DNA sequences and RBBHs using protein sequences. This dual prediction approach allows POB to address three common drawbacks encountered by RBBH-based orthology prediction methods: (i) it alleviates incompleteness or base–error problems (deletions, inserts, or sequencing

errors) present in some protein sequences found in public databases, which are typically used in orthology prediction because of their higher conservation when compared with equivalent DNA sequences; (ii) it minimizes the number of false positives caused by sometimes inconsistent hits ranking produced by local sequence alignment tools such as BLAST; and (iii) it ensures that sequence similarities obtained via local sequence alignment algorithms using DNA and protein specific similarity matrices and other related parameters are consistent. The POB also offers robust free-text search functionalities with recommendations and auto completion that enable users to navigate and filter orthology and functional gene information. Interspecies visual comparisons offered by POB consist of intricate gene-order representations and rather unique interactive synteny plots (first interactive syntenic dot plots to our best knowledge), allowing users to explore and pinpoint individual gene impact and location on each plot. Unlike other orthology database repositories, such as PGDD (Lee et al., 2013) and Plaza (Proost et al., 2015), POB uses existing gene identifiers rather than introducing new or modifying existing ones, thus minimizing the need of cross-identification mapping tools. Moreover, POB provides users with an application program interface (API) that facilitates integration with other tools and a downloadable standalone version for Microsoft Windows for convenient use on a personal computer.

Materials and Methods

Data Sources

We downloaded 753,702 DNA coding sequences, the corresponding protein sequences, physical maps, and GO vocabulary for functional annotation from the file transfer protocol (FTP) site of Ensembl Plants release 33 (Monaco et al., 2014) for 20 plant species: *Aegilops tauschii* Coss., *Arabidopsis thaliana* (L.) Heynh., sugar beet (*Beta vulgaris* L.), *Brachypodium distachyon* (L.) Beauv., turnip (*Brassica rapa* L.), soybean [*Glycine max* (L.) Merr.], barley (*Hordeum vulgare* L.), barrel clover (*Medicago truncatula* Gaertn.), banana (*Musa acuminate* Colla), rice (*Oryza sativa* L. subsp. *japonica* Kato), *Ostreococcus lucimarinus,* foxtail millet [*Setaria italica* (L.) P. Beauv. *subsp. italica*], tomato (*Solanum lycopersicum* L.), potato (*Solanum tuberosum* L.), sorghum (*Sorghum bicolor* L.), red clover (*Trifolium pratense* L.), wheat (*Triticum aestivum* L.), red wild einkorn (*Triticum urartu* Tumanian ex Gandilyan), grapevine (*Vitis vinifera* L.), and maize (*Zea mays* L.) (Fig. 2; Table 1). In two cases (wheat and maize) we used sequences from Ensembl Plants release 32, since the newly adopted accession numbers for the two species do not match any existing public databases other than the source of origin. The plant species have been carefully selected to satisfy at least one of the two criteria: (i) model organisms with well-annotated genes and proteins and (ii) plant species evolutionary closely related to wheat.
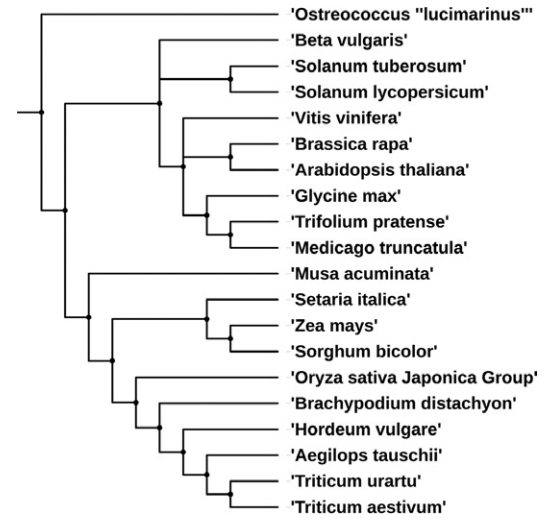

Fig. 2. The phylogenetic tree for the 20 plant species supported in Plant Orthology Browser. The figure was obtained with the Interactive Tree of Life (iTOL) v2 (Letunic and Bork, 2011) and the NCBI Taxonomy Common Tree (Federhen, 2012).

Coding DNA and protein sequences were preprocessed and only those corresponding to the longest transcripts were selected for pairwise BLAST runs. Annotations and physical map information was acquired programmatically from the Gramene MySQL database build 51. The physical mapping information for *A. tauschii* and red wild einkorn was complemented with information extracted and processed from the original publications of the two species (Jia et al., 2013; Ling et al., 2013).

System Architecture

The POB system includes two major components: (i) the orthology prediction and data acquisition pipeline and (ii) the POB web interface powered by an in-memory no-SQL back end.

The orthology prediction and data acquisition pipeline were implemented in Perl, Go (https://golang.org/), and C and consists of two modules (top of Fig. 3). The first module is the data acquisition and update module, which consists of a series of custom built Perl scripts that connect automatically to external reference public data repositories such as Ensembl Plants, Gramene, and the Gene Ontology Consortium and extract sequences, physical maps, and annotations for all 20 plant species supported by POB. The data is curated and preformatted to be fed into the downstream modules for further processing. The second module is the orthology prediction and postprocessing module. This module includes custom-built Perl and C orthology prediction software. The module takes as input a complete set of plant genome DNA and protein sequences and predicts a set of putative high-confidence orthologs for every pair of plant species that represent the overlap of DNA and protein based RBBHs. The RBBHs are obtained using *blastn* (for DNA sequences) and *blastp* (for protein sequences) pairwise similarity scores calculated in a similar fashion as InParanoid (Ostlund et al., 2010) does for

Table 1. Twenty plant species supported by Plant Orthology Browser and their characteristics extracted from Ensembl Plants (release 33) and publications.

No.	Plant species	No. genes	Classification	Ensembl Plants assembly	European Nucleotide Archive (ENA) assembly ID	Genome size (Mbp)	Reference
1	*A. tauschii*	33,849	Monocotyledons	ASM34733v1	GCA_000347335.1	4230	Jia et al., 2013
2	*A. thaliana*	32,243	Eudicotyledons	TAIR10	GCA_000001735.1	135	Arabidopsis Genome Initiative, 2000
3	*B. vulgaris*	27,421	Eudicotyledons	RefBeet-1.2.2	GCA_000397105.1	758	Dohm et al., 2013
4	*B. distachyon*	26,552	Monocotyledons	v1.0	GCA_000005505.1	355	Ozdemir et al., 2008
5	*B. rapa*	41,018	Eudicotyledons	IVFCAASv1	GCA_000309985.1	529	Johnston et al., 2005
6	*G. max*	46,430	Eudicotyledons	V1.0	GCA_000004515.1	1115	Schmutz et al., 2010
7	*H. vulgare*	24,287	Monocotyledons	ASM32608v1	GCA_000326085.1	5300	Wicker et al., 2004
8	*M. acuminata*	36,542	Monocotyledons	MA1	GCA_000313855.1	523	D'Hont et al., 2012
9	*M. truncatula*	50,894	Eudicotyledons	MedtrA17_4.0	GCA_000219495.1	384	Tang et al., 2014
10	*O. lucimarinus*	7651	Monocotyledons	ASM9206v1	GCA_000092065.1	13.2	Palenik et al., 2007
11	*O. sativa* L. subsp. *japonica*	35,679	Monocotyledons	IRGSP-1.0	GCA_001433935.1	420	Goff et al., 2002
12	*S. italica*	35,471	Monocotyledons	JGIv2.0	GCA_000263155.1	405	Bennetzen et al., 2012; Zhang et al., 2012
13	*S. lycopersicum*	34,727	Eudicotyledons	SL2.50	GCA_000181095.1	900	Tomato Genome Consortium, 2012
14	*S. tuberosum*	39,031	Eudicotyledons	SolTub_3.0	GCA_000226075.1	844	Xu et al., 2011
15	*S. bicolor*	34,496	Monocotyledons	Sorbi1	GCA_000003195.1	730[15]	Paterson et al., 2009
16	*T. pratense*	40,868	Eudicotyledons	Trpr	GCA_000583005.1	420	De Vega et al., 2015
17	*T. aestivum*	102,235	Monocotyledons	IWGSC1+popseq	GCA_000188135.1	17,000	International Wheat Genome Sequencing Consortium, 2014
18	*T. urartu*	34,903	Monocotyledons	ASM34745v1	GCA_000347455.1	3920	Ling et al., 2013
19	*V. vinifera*	30,434	Eudicotyledons	IGGP_12x	GCA_000003745.1	487	Jaillon et al., 2007
20	*Z. mays*	38,735	Monocotyledons	AGPv3	GCA_000005005.5	2400	Schnable et al., 2009
TOTAL		403,997					

protein sequences. For each pair of plant species (P1, P2) and each type of sequence (DNA, protein), four BLAST runs are executed: P1 against P2, P2 against P1, P1 against P1, and P2 against P2. The first two BLAST output files are used for orthology prediction, while the last two are used for the detection of in-paralog clusters. The difference between our approach and InParanoid consists in the way orthologs are defined. In POB, orthologs denote all RBBHs typically included in a singleton cluster in InParanoid and the most relevant member (the in-paralog with the highest similarity) of the interspecies in-paralog clusters. The inclusion of a single in-paralog representative (with highest similarity) from a cluster of one-to-many or many-to-many in-paralog cluster is justified by two factors: (i) to avoid the complete blank out of entire chromosomes from partial or entire genome duplications, and (ii) to keep the number of false positives to a minimum based on previous observations reported by InParanoid authors (Remm, et al., 2001). To ensure that POB is not hampered by a major loss of orthologs because of the aforementioned definition, we calculated the overlap between the number of POB orthologs and Ensembl Plants orthologs (one-to-one, one-to-many, many-to-many) for six pairs of species: *B. distachyon* vs. *A. tauschii*, *A. thaliana*, barley, *O. sativa* subsp. *japonica*, sorghum, and red wild einkorn. The number of commonly predicted orthologs (Table 2) ranged between

86 and 94%, which we deem as being acceptable in the absence of a gold standard for plant orthology.

This approach is crucial for plant genomes given the high degree of sequence duplications and polyploidy, which typically leads to high numbers of false positives (paralogs). Therefore, we acknowledge that some one-to-many and many-to-many orthology relationships are absent in POB. We also acknowledge that currently POB does not include homeologs defined in (Glover, et al., 2016) as same species genes originated by a speciation event, which were brought back together in the same genome by allopolyploidization, but we plan to address these insufficiencies in future releases.

The BLAST runs are performed in parallel for each sequence type and provide up to four times speed up compared with the sequential alternative. Each BLAST run uses the following parameter settings: *-m* 8, *-W* 9 (for DNA) or 3 (for proteins) and *-e* 1×10^{-4}. The E-value cutoff (1×10^{-4}) is applied to avoid short domain-level matches.

A total of 1,021,611 orthologs (Fig. 4) have been predicted in total among all possible pairs of plant species considered here. Once the orthologs are predicted, physical mapping information is acquired from Ensembl BioMart. Using this information, relative gene order for all genes in all chromosomes for all plant species is calculated using a Go script, and it is further used for synteny calculations using a Go language implementation of
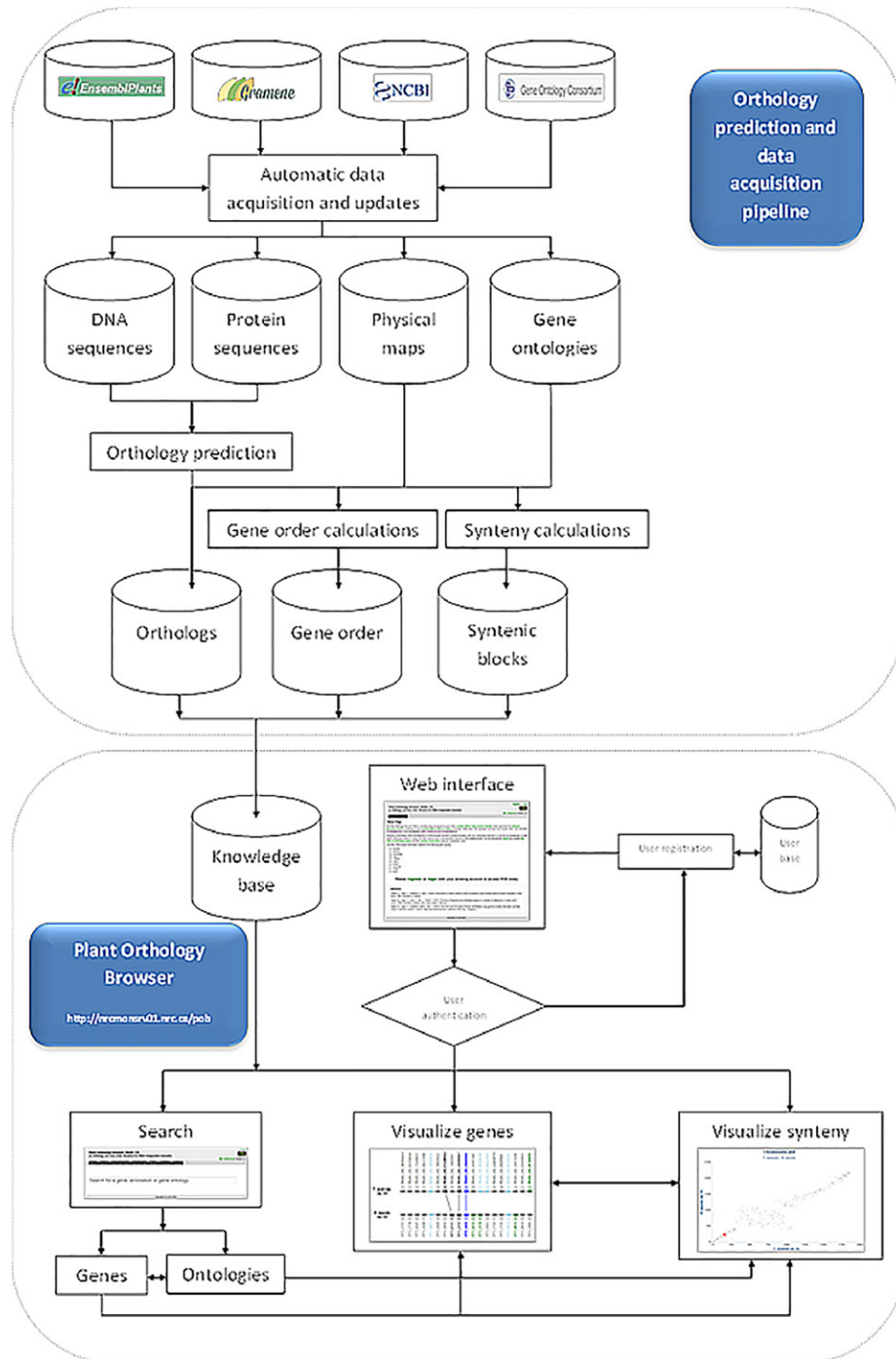
Fig. 3. Layout of the main Plant Orthology Browser (POB) components. The two main components of POB are (i) the orthology prediction and data acquisition pipeline and (ii) the POB (http://nrcmonsrv01.nrc.ca/pob/).

DAGchainer (Haas et al., 2004). DAGchainer accepts as input a file with orthologs between two species, the corresponding match E-values and their positions on the chromosomes. It then starts to calculate chains of gene pairs sharing conserved order between genomic regions by identifying the highest scoring paths in a directed acyclic graph using a dynamic programming approach. The minimum size of a block is three (i.e., three orthologs) and the maximum distance between two chains (conserved blocks) is 200 kb. Additionally, we use a gap extension penalty equal to –3, a gap length of 10 kb, minimum

alignment score equal to 150, and a gap open penalty equal to 0. Reverse blocks are identified by running DAG-chainer on the reversed version of one of the two chromosomes.

The POB web interface currently implements three interconnected modules: the gene-order visualization module, the synteny visualization module, and the search module.

The gene-order visualization module offers users a highly interactive environment for exploration of gene order between any pair of chromosomes in any

Table 2. Number of commonly predicted orthologs between Ensembl Plants (release 33) and Plant Orthology Browser (POB) for six pairs of species.

| | POB orthologs | Ensembl Plants | | | |
		1-to-1 orthologs	1-to-many orthologs	Many-to-many orthologs	Total overlap
B. distachyon vs. *A. tauschii*	14,934	10,980 (73.5%)	2339 (15.7%)	498 (3.3%)	13,817 (92.5%)
B. distachyon vs. *A. thaliana*	10,894	4961 (45.5%)	2926 (26.9%)	2104 (19.3%)	9991 (91.7%)
B. distachyon vs. *H. vulgare*	14,789	12,005 (81.2%)	1512 (10.2%)	222 (1.5%)	13,739 (92.9%)
B. distachyon vs. *O. sativa japonica*	16,525	13,141 (79.5%)	1910 (11.6%)	407 (2.5%)	15,458 (93.5%)
B. distachyon vs. *S. bicolor*	17,415	13,553 (77.8%)	2283 (13.1%)	548 (3.1%)	16,384 (94.0%)
B. distachyon vs. *T. urartu*	15,798	10,851 (73.9%)	2296 (15.6%)	437 (3.0%)	13,584 (86.0%)

	A. tauschii	A. thaliana	B. distachyon	B. rapa	H. vulgare	O. sativa	S. bicolor	T. aestivum	T. urartu	Z. mays
A. tauschii		2760	13026	2394	12610	10713	11402	14830	16065	9552
A. thaliana			3594	15383	2771	3241	3570	1551	2765	2879
B. distachyon				3078	13331	14384	15309	11050	12798	12724
B. rapa					2373	2763	3016	1365	2322	2516
H. vulgare						10992	11741	10573	12132	9947
O. sativa							14443	8719	10640	12061
S. bicolor								9051	11183	16164
T. aestivum									13181	7658
T. urartu										9291
Z. mays										

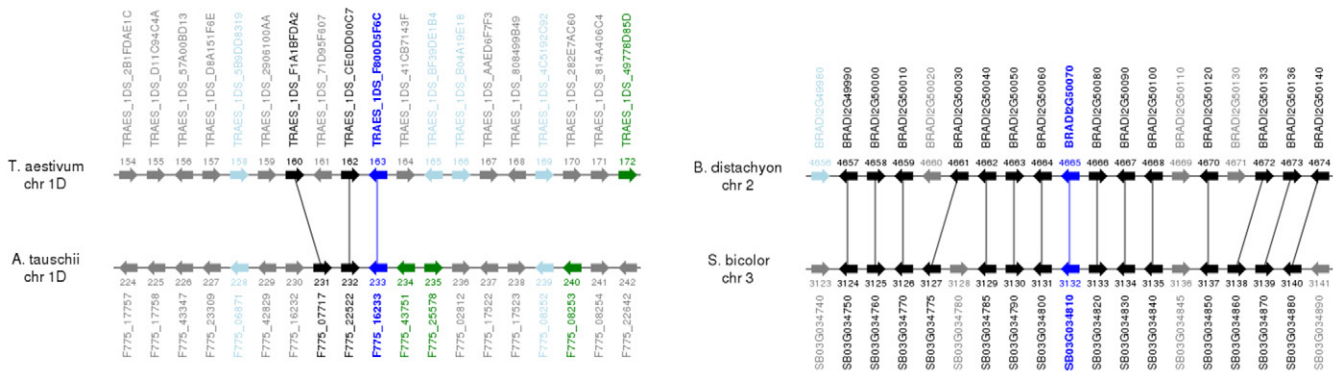Fig. 4. Total number of orthologs predicted by OrthoPred for each pair of plant species.

Fig. 5. Gene order view for a selected ortholog (depicted in blue) surrounded by neighboring genes. Left: weak conservation of local gene order between genes on *T. aestivum* chromosome 1D and *A. tauschii* chromosome 1D. Right: strong conservation of local gene order between genes on *B. distachyon* chromosome 2 and *S. bicolor* chromosome 3.

plant species supported in POB (Fig. 5). Plant Orthology Browser provides agile browsing of orthologous pairs of genes combined with one-click connections to external data sources where gene cards are available. A what-you-see-is-what-you-get (WYSIWYG) approach is implemented, and everything can be exported as either a publication-quality image (GIF, JPG, PNG, or SVG) or text depending on the nature of information that is required. A one-click transition between this module and the synteny visualization module is also provided.

The synteny visualization module provides a "bird's-eye view" of the full extent of orthology relations between

a pair of chromosomes in two distinct plant species (Fig. 6). Using an interactive dot plot graph implemented using JQuery and HighCharts (http://www.highcharts.com/), the user can select one or multiple orthologs, zoom in and zoom out of areas of interest, and switch between three subviews that include intuitive highlighting for genes, conserved syntenic blocks (Fig. 7), and ontologies. The same WYSIWYG approach is applied here, too, and a one-click transition between this module and the previous one is provided based on the selected ortholog of interest.

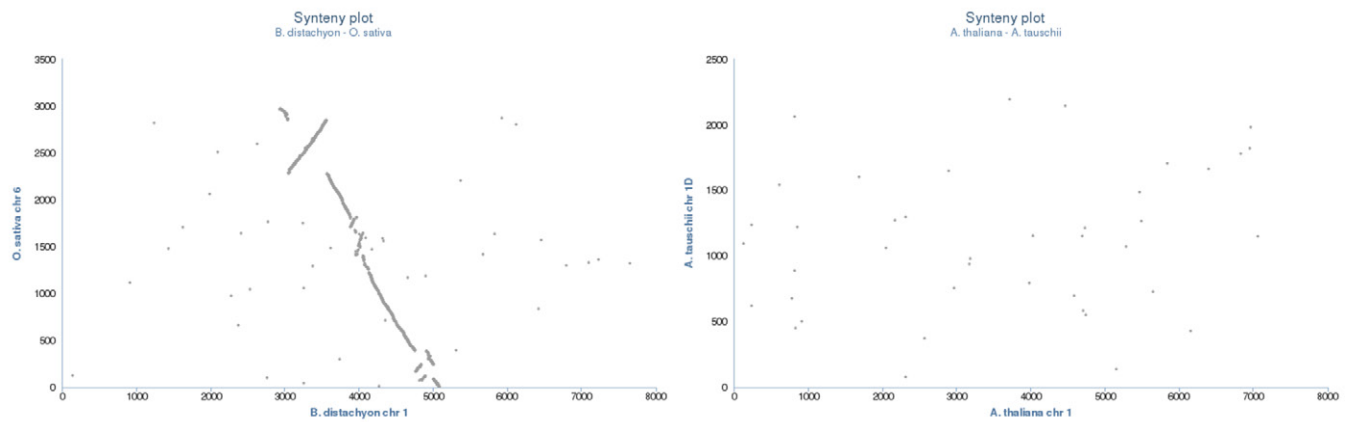The search module realizes the interconnectivity among all modules via a customized free-text search

Fig. 6. Synteny views. Left: conserved synteny between *B. distachyon* chromosome 1 and *O. sativa* chromosome 6. Right: no syntenic conservation between *A. thaliana* chromosome 1 and *A. tauschii* chromosome 1D.
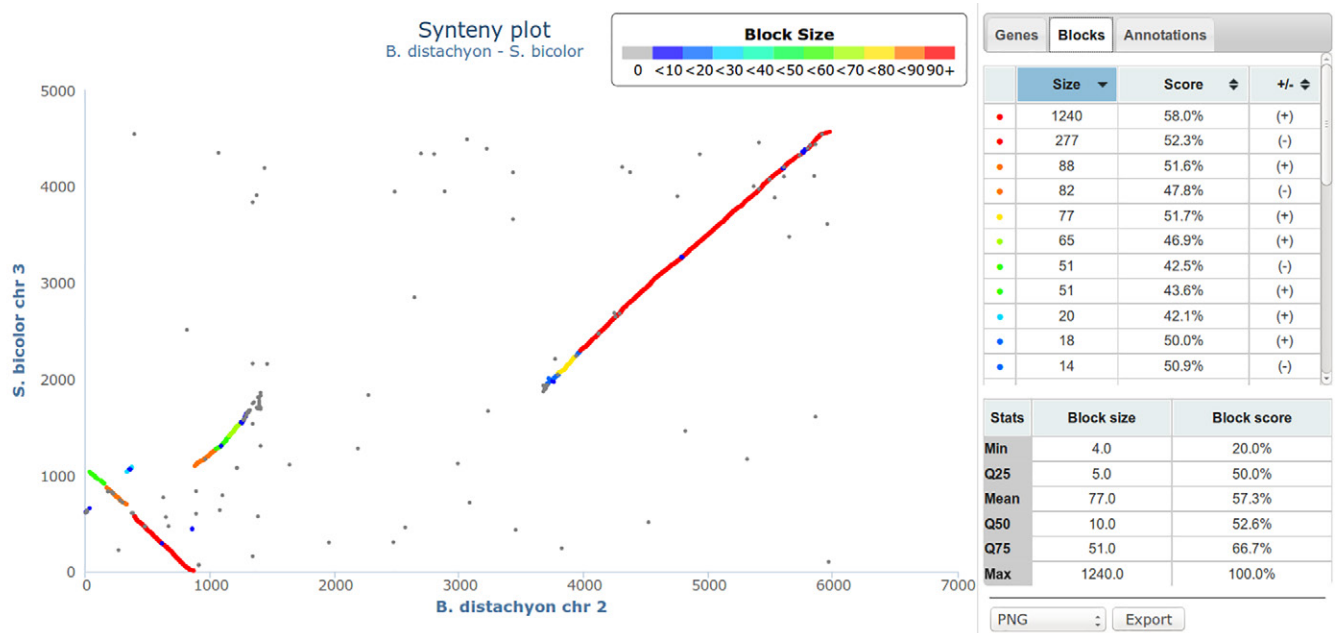

Fig. 7. Highlighted blocks in synteny view of *S. bicolor* chromosome 3 and *B. distachyon* chromosome 2. Left side of the image provides a dot plot visualization of the conserved syntenic blocks in chromosome 3 of sorghum and chromosome 2 of Brachypodium. The top right table provides access to score size and density scores for each conserved syntenic block. The bottom right table provides basic statistics (minimum, Q25, median = Q50, mean, Q75 and maximum) for all conserved syntenic blocks depicted in the dot plot. All the information (figure and tabular data) can be exported as a publication-quality image or text files.

capability with online as-you-type suggestions and highlighting that allows users to explore various facets of the underlining information (Fig. 8) without being constrained by interface-dependent search fields.

Based on the user input, the output is divided into two categories: gene sequence and gene ontology information. The user can easily explore both types of information by flipping between two panels (top right and bottom right images in Fig. 8) with links conveniently placed at the top of the visualization area.

The underlining mechanism behind the search module is a suffix array, which includes orthologous gene pairs, gene annotations such as synonyms (gene names, Affymetrix IDs), gene ontologies (IDs, names, definitions), and gene physical mapping information.

The suffix array is prebuilt using Go language and loaded in memory, thus conferring higher information access speeds (up to 10 times faster) than database- or file-based implementations. The information in the suffix array is grouped by species. The suffix array search results are further mapped and combined with gene-order information accessed from text files, thus making a compromise between higher speeds and oversized memory objects. Overall, the in-memory search suffix array provides very fast search results, which are within the 12-s user-patience limit for website responsivity.

On top of these modules, POB provides users with means to retrieve subsets of orthologs for a given set of gene identifiers from any of the 20 currently supported plant species. The *find-my-orthologs* feature implements
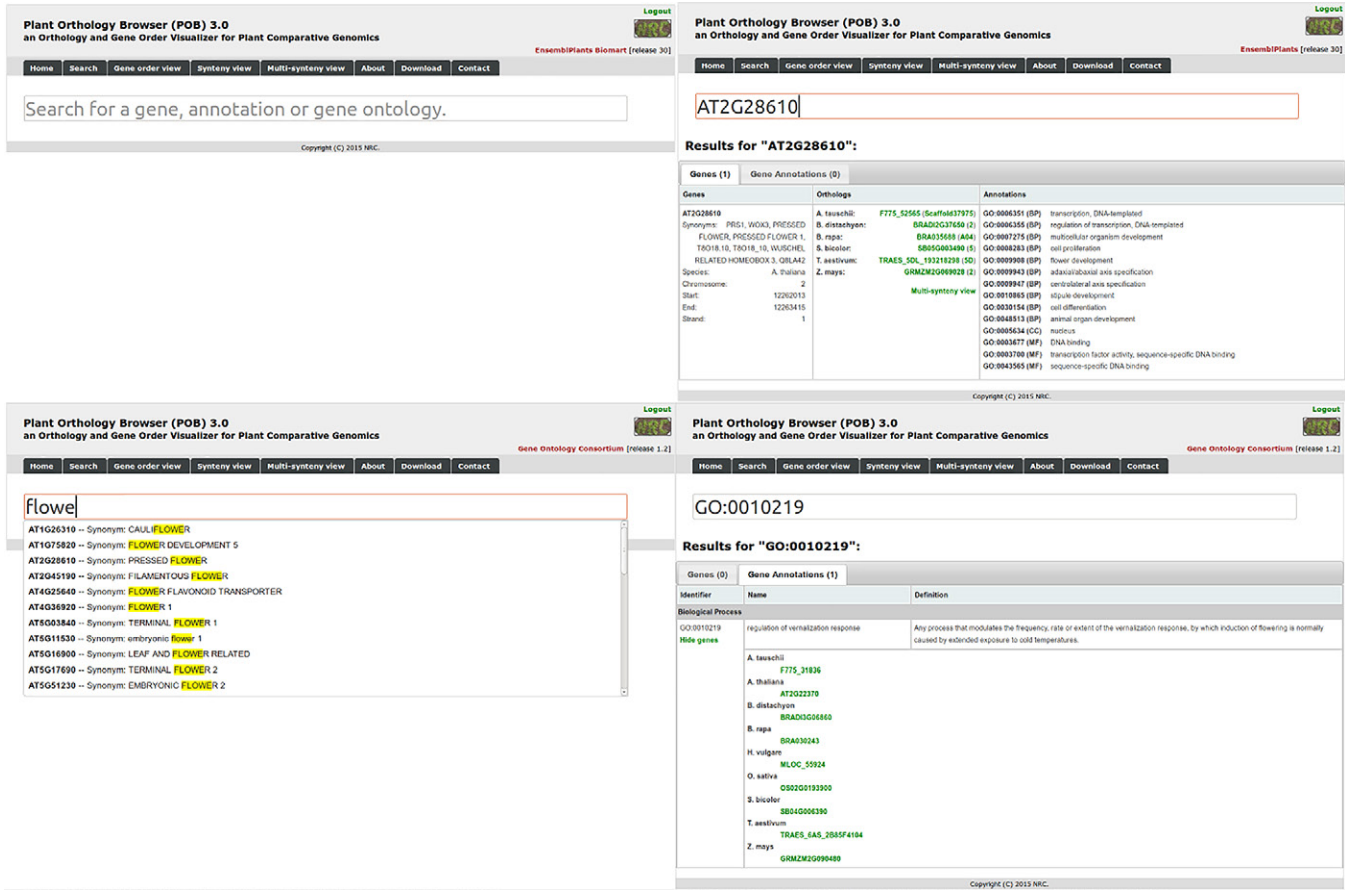
Fig. 8. The search module implemented in Plant Orthology Browser (POB). Top left figure depicts the main page of the search module. The user can provide a complete or partial gene ID, annotation keywords, or gene ontology (GO term or keywords). Once the user starts typing, suggestions of potential matches of interest are displayed (bottom left view). If the user provides a complete or incomplete gene ID (top right figure), then POB will display tabular information including gene matches (Ensembl id, synonyms, species name, chromosome localization, start and end, and strand orientation), list of orthologs with direct links to the other two visualization modules, and GO annotations for each matched gene. If the user provides a complete or incomplete GO term, POB will output a tabular view including matching GO terms (and a list of genes in all species that contain that GO term with links to the other visualization modules) and the corresponding GO term names and definitions.

such functionality by means of an API capable to support more complex user requirements (e.g., programmatic retrieval of full sets of orthologs) and described in the sections below. The multisynteny view functionality provides user with simultaneous access to synteny views of all orthologs for a gene in a given species.

Orthology Prediction

We implemented a RBBH approach for one-to-one orthology prediction inspired from previous work applied to human and mouse genomes (Lynn et al., 2008). Two BLAST runs are executed for each pair of plant species and for each sequence type (DNA, protein) to identify RBBHs. An orthology relationship is assigned for those pairs of genes that are bidirectional hits within a confidence interval (E-value $\leq 10^{-4}$), which includes all singleton RBBHs and the most relevant in-paralogs as described in detail in the previous sections. Complete lists of orthologs for each pair of species are available for download from the POB website: http://nrcmonsrv01.nrc.ca/pob/about.html. The POB orthology detection approach is similar with the one

implemented in InParanoid for protein sequences, and we refer the readers for details to the original manuscript (Ostlund et al., 2010). As it has been already reported previously (Despalins et al., 2011), the need for high computational power and considerable computing-time requirements (up to 4 h for large genomes such as that of wheat), makes any online orthology prediction calculations with user-uploaded plant genome data on the POB website impractical. Therefore all the POB calculations related to orthology and synteny prediction are performed offline. Nevertheless, we invite users to suggest plant species that are desired to be integrated into POB and we will kindly add them to the repository.

Application Program Interface, Customization, and Standalone Version

The POB provides basic API-based interactions for users who need to use its functionality programmatically. For example, full lists of orthologs between any two supported plant species can be acquired by providing the two species scientific names, where the first letter is capitalized

and an underscore character is used as word separator. For example, the list of orthologs for *B. distachyon* and sorghum can be obtained using the following API call: http://nrcmonsrv01.nrc.ca/pob/api/index.php/rbbh/orthopred:intersection/Brachypodium_distachyon/Sorghum_bicolor. Similarly, complete lists of orthologs for one query gene can be obtained by calling http://nrcmonsrv01.nrc.ca/pob/api/index.php/rbbh_by_gene/SpeciesID/orthopred:intersection/GeneID and replacing SpeciesID and GeneID with the corresponding Ensembl Plants species name (e.g., *Triticum_aestivum*) and gene identifier (e.g., *TRAES3BF000100010CFD_G*), respectively.

Because of the clear separation between supporting information and web functionality, the POB can be customized to host information for nonplant species (e.g., mammals, fish, fungi, etc.) or include orthologs obtained with other methods (e.g., phylogenetic) with minimum implementation efforts. For convenience, we also provide users with a Microsoft Windows standalone version of the POB that offers the same functionality as the online one and can be used on a personal computer.

Results

This work is focused on the comparative genomics analyses of orthologs in plants with medium and large size genomes. Nine eudicotyledon and 11 monocotyledons plant species are currently supported by POB (Table 1). In the following sections, we describe the main functionalities of the browser and their utility and usefulness for computational biologists.

The Gene Order View

This page enables close-up orthology visualization of gene pairs lying on pairs of chromosomes in two plant species. It contains a form where users can chose two plant species, the type of orthology prediction (only DNA, only protein, intersection) and a query Ensembl Plants gene ID. The results are displayed as an interactive image in a fixed-size window (19 genes long) centered on the main orthologous gene pair (the top one is the query gene) colored in blue. The two chromosomes are represented by two straight black lines. Each gene is represented by an arrow and each orthology relationship consists of a line that connects two arrows. The direction of the arrow coincides with the strand direction. We use different colors for arrows to depict four distinct scenarios. Blue arrows represent the query gene and its ortholog. Black arrows represent genes whose orthologs fall within the visualization window. Light blue arrows are genes whose orthologs fall on the same chromosomes but outside the visualization window. Green arrows are genes whose orthologs fall on a different chromosome. Gray arrows represent genes that do not have a predicted ortholog in POB. Each gene is labeled with its corresponding Ensembl Plant ID and its relative position counting from the start of the chromosome. The image is interactive and allows multiple types of interactions. The chromosomes are mobile and can be dragged either together or one at a time to the left or to the right for better positioning of the genes of interest. A simple sideways mouse drag allows for moving one gene at a time. If CTRL or SHIFT keys are pressed while dragging, faster movement is performed to either one window size (19 genes) or 100 genes at a time, respectively. The black, light blue and green gene symbols can be clicked on to center the view on the selected ortholog. Gene IDs are directly connected to corresponding Ensembl Plants summary pages. All graphical results can be downloaded as SVG, JPG, GIF and PNG images. The page also provides a transition to the synteny visualization of all orthologs on a pair of chromosomes in two plant species.

The Synteny View

Apart from the gene-order view, POB also enables synteny visualization of complete sets of orthologs on two chromosomes in two plant species. The synteny view page consists of three parts: (i) a form where the user can select the two plant species, the type of orthology prediction, and the chromosomes in each plant; (ii) a plotting area, which includes an interactive dot plot representation of orthologs; and (iii) an interactive information area with three tabs (genes, blocks, annotations), which allows further exploration and visualization of orthologous genes with specific features.

The synteny plot consists of an interactive dot plot, with *x*- and *y*-axis representing gene relative positions on two chromosomes in two plants. Each dot represents an orthologous pair of genes with relative positions corresponding to the dot projection coordinates on the *x*- and *y*-axis (Fig. 6). The user can click on each dot (pair of orthologous genes), which will add an ortholog card in the genes tab of the information panel on the right. By holding CTRL, the user can select more orthologs at once. The ortholog card includes the gene IDs of the two orthologous genes, their strand orientation (+ or −), block information (if the ortholog belongs to a predicted conserved block), and GO information for the two genes. If block information is provided, hovering over it will highlight the synteny block in the dot plot where the orthologous gene pair resides. The GO information is partitioned in three sections. First section includes GO terms common to both genes, while the following two sections describe GO terms specific to each of the two genes.

The genes tab provides the user with gene search capabilities, which allows for exact localization of the orthologous pair including the gene of interest in the dot plot. Once a gene ID is provided, the resulting ortholog that contains the query gene ID will be highlighted with black in the dot plot representation. To further enhance the exploration capabilities of the dot plot, the user can zoom in and zoom out.

Switching on the blocks tab will enable a colorful representation of all synteny blocks in the dot plot. The conserved synteny blocks are colored based on their length, from dark blue, for blocks with up to nine orthologs, to

intense red for blocks with more than 90 orthologs. The top side of the information panel contains detailed information about block sizes, calculated block scores based on the block density (Eq. [1]), and block orientation.

$$\text{Block score} = \frac{\text{No. orthologous block genes}}{\text{Total number of block genes}} \times 100 \quad [1]$$

Hovering over each line with block information will highlight the complete block in the dot plot representation. Clicking on each line will select it and maintain the highlight in the dot plot. The bottom side contains a tabular representation of descriptive statistics for block sizes and scores for the orthologs on the selected two chromosomes.

The annotations tab provides yet another level of exploration, by highlighting orthologs characterized by similar GO annotations. The user can search for a specific GO term ID, GO term name, or annotation keywords. Autocompletion suggestions are provided, which include a list of GO term ids, their name, and the number of orthologs that will be highlighted in the dot plot. Once GO terms are selected, the highlighted orthologs (dots) will change their color based on the number of annotations corresponding to the two orthologous genes. A similar color scheme as the one used for blocks highlighting is employed, ranging from blue shades for low frequency annotated genes (up to four genes) to red for highly annotated genes (>10 genes). A GO term card will appear in the information box underneath the search field, which includes the GO term ID, the number of genes characterized by this term in the current visualization plot, the name, namespace and definition of the GO term, and the number of genes in the plant species from where the annotation originated.

Keyword Search

For situation where users do not have any prior information, such as a gene identifier or a set of genes of interest, the search module allows an unconstrained keyword search backed up by a powerful autocompletion suggestion engine with highlighting. The user can use part of a gene name, gene ID, or GO term annotation to prime their search. The search returns potential matches that fall within two categories included in two separate tabs: genes and gene annotations.

The gene results include genes and their corresponding orthologs and annotations. The matching is performed based on gene IDs and synonyms that contain or are identical with the query string. When available, lists of orthologs are provided for each plant species and each orthologous gene and corresponding chromosomes are linked directly to the gene-order and synteny-view modules, respectively. The annotations include lists of GO terms corresponding to each matching gene ID, their definition, and the major category they belong to (biological process, cellular component, or molecular function).

The gene annotations results provide a complementary matching capability, where the query keyword matches the GO term ID, the GO term name, or a part

of its definition. The search results are presented in a tabular format, which includes the three fields used for matching. For each matching GO term ID, we retrieve the gene IDs that have that term as part of their annotation. An expandable link containing the total number of genes with this property is provided for user interaction. Once the user interacts with the link, a complete list of gene IDs and corresponding plant species is provided, each gene ID being connected to a backward search leading to detailed information about it.

Discussion

The construction and development of the POB was undertaken to enable the plant research community to carry out generic and targeted comparative genomics-based exploration of orthologous genes among well-studied model organisms, such as *A. thaliana*, and recently sequenced nuclear cereal genomes such as those of *A. tauschii*, red wild einkorn, and wheat, including visualization of gene order, synteny conservation and functional annotations of the underlying DNA and protein sequences. The POB offers various visualization modules for orthology-based comparative genomics studies of 20 plant species, and it is the only plant-based resource offering highly interactive and interconnected dot plot and chromosome-pairs visual representations of gene locations and gene functional annotations on pairs of chromosomes. Furthermore, the synteny view within the POB is a unique plant orthology tool that offers interactive combined visualization of syntenic gene blocks and gene functional annotations, revealing colocalization or spread of functionally equivalent genes on chromosomes. The POB results can thus provide opportunities for deeper exploration of gene-order conservation among recently sequenced cereals and model organisms. Another unique feature of the POB is the availability of overlapping predictions for DNA- and protein-based orthologs, which increase the level of prediction confidence while reducing the number of false positives. The complete sets of orthologs hosted by the POB was successfully used to enrich gene annotations via orthology cliques and phylogenetic scoring of distinct GO terms that characterize various plant species in previous research projects (Tulpan et al., 2015). Most of the existing plant orthology browsers, such as GMOD-GBrowse_syn (http://gmod.org/wiki/GBrowse_syn), SynView (http://eupathdb.org/apps/SynView/), and Sybil (http://sybil.sourceforge.net/), provide conventional orthology information, limited or no selection of plant species, and static images or limited interactivity views for chromosome synteny and conserved gene order. Table 3 presents a summary of features selected based on their practical relevance that show how the POB stands apart from similar solutions, and it is prone to be a serious contender in the already fragmented and diluted bioinformatics software space.

In summary, we believe that POB brings a different level of interactivity, multilevel information accessibility, and visualization among the plant orthology platforms,

Table 3. Summary of features available in Plant Orthology Browser (POB), GBrowse_syn, SynView, and Sybil.

Feature	POB	GBrowse_syn	SynView	Sybil
Synteny view	Interactive	Interactive	Interactive	Interactive
Dot–plot view (full chromosomes)	Interactive	None	None	Uses synteny gradient view instead
Gene annotations	Interactive	Fixed format	None	None
Built-in search	Free-text	Fixed format	None	None
Scalability	High	High	High	Low
Easiness to adapt to other data types	High	High	Low	Low
Easiness to add new features	High	Medium	Medium	Low
Support for multiple species	Yes	Yes	Yes	Only bacteria
Applied to large data sets (e.g. plants)	Yes	Yes	No	No
Support API for batch data queries	Yes	Yes	No	No
Embedded web server	Yes	No	No	No

and therefore, it will be well received by the plant biology research community. The innovative visualization, data organization, and information accessibility approaches implemented in the POB would open the path for future research initiatives on biological data visualization and interpretation. Online documentation and information are also available within the POB for understanding the resource and its capabilities.

Conclusions

The POB is a highly interactive, scalable, and customizable web resource for specialized visual inspection of orthology among genes from 20 different plant species. It is freely available (http://nrcmonsrv01.nrc.ca/pob), and it is setup to be automatically updated with the latest version of plant gene models, physical maps, and annotations from Ensembl Plants, BioMart, Gramene, and the Gene Ontology Consortium. In future versions, we plan to add new features and functionality to the browser, of interest for the plant biology research community, such as multispecies genome-level orthology visualization, polyploidy and homeology views and integration of biochemical pathways information. Thanks to POB's easily customizable data content, wheat-centric and chloroplast-based instances are under development and will be published later this year. Furthermore, with the progress of genome sequencing efforts for new wheat-related plant species, such as rye (*Secale cereale* L.) and the rye–wheat hybrid, triticale (×*Triticosecale* spp.), we hope to integrate their genome information into the POB to enable a more complete global visualization and analyses. We strongly encourage and welcome constructive suggestions for existing and new features from the plant genome research community.

References

Altenhoff, A.M., and C. Dessimoz. 2009. Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods. PLoS Comput. Biol. 5:e1000262. doi:10.1371/journal.pcbi.1000262

Altenhoff, A.M., A. Schneider, G.H. Gonnet, and C. Dessimoz. 2011. OMA 2011: Orthology inference among 1000 complete genomes. Nucleic Acids Res. 39:D289–D294. doi:10.1093/nar/gkq1238

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815. doi:10.1038/35048692

Bennetzen, J.L., J. Schmutz, H. Wang, R. Percifield, J. Hawkins, A.C. Pontaroli, et al. 2012. Reference genome sequence of the model plant *Setaria*. Nature Biotech. 30:555–561. doi:10.1038/nbt.2196

Chen, F., A.J. Mackey, C.J. Stoeckert, and D.S. Roos. 2006. OrthoMCL-DB: Querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res. 34:D363–D368. doi:10.1093/nar/gkj123

Curtis, D.S., A.R. Phillips, S.J. Callister, S. Conlan, and L.A. McCue. 2013. SPOCS: Software for predicting and visualizing orthology/paralogy relationships among genomes. Bioinformatics 29:2641–2642. doi:10.1093/bioinformatics/btt454

Dalquen, D.A., and C. Dessimoz. 2013. Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. Genome Biol. Evol. 5:1800–1806. doi:10.1093/gbe/evt132

Deluca, T.F., I.H. Wu, J. Pu, T. Monaghan, L. Peshkin, S. Singh, and D.P. Wall. 2006. Roundup: A multi-genome repository of orthologs and evolutionary distances. Bioinformatics 22:2044–2046. doi:10.1093/bioinformatics/btl286

Despalins, A., S. Marsit, and J. Oberto. 2011. Absynte: A web tool to analyze the evolution of orthologous archaeal and bacterial gene clusters. Bioinformatics 27:2905–2906. doi:10.1093/bioinformatics/btr473

De Vega, J.J., S. Ayling, M. Hegarty, D. Kudrna, J.L. Goicoechea, Å. Ergon, et al. 2015. Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. Sci. Rep. 5:17394. doi:10.1038/srep17394

D'Hont, A., F. Denoeud, J.M. Aury, F.C. Baurens, F. Carreel, O. Garsmeur, et al. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature 488:213–217. doi:10.1038/nature11241

Dohm, J.C., A.E. Minoche, D. Holtgräwe, S. Capella-Gutiérrez, F. Zakrzewski, H. Tafer, et al. 2013. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature 505:546–549. doi:10.1038/nature12817

Fang, G., N. Bhardwaj, R. Robilotto, and M.B. Gerstein. 2010. Getting started in gene orthology and functional analysis. PLoS Comput. Biol. 6:e1000703. doi:10.1371/journal.pcbi.1000703

Federhen, S. 2012. The NCBI taxonomy database. Nucleic Acids Res. 40:D136–D143. doi:10.1093/nar/gkr1178

Flicek, P., M.R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, et al. 2014. Ensembl 2014. Nucleic Acids Res. 42:D749–D755. doi:10.1093/nar/gkt1196

Fulton, D.L., Y.Y. Li, M.R. Laird, B.G.S. Horsman, F.M. Roche, F.S.L. Brinkman. 2006. Improving the specificity of high-throughput ortholog prediction. BMC Bioinf. 7:270. doi:10.1186/1471-2105-7-270

Glover, N.M., H. Redestig, and C. Dessimoz. 2016. Homoeologs: What are they and how do we infer them? Trends Plant Sci. 21:609–621. doi:10.1016/j.tplants.2016.02.005

Goff, S.A., D. Ricke, T.H. Lan, G. Presting, R. Wang, M. Dunn, et al. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92–100.

Haas, B.J., A.L. Delcher, J.R. Wortman, and S.L. Salzberg. 2004. DAG-chainer: A Tool for mining segmental genome duplications and synteny. Bioinformatics 20:3643–3646. doi:10.1093/bioinformatics/bth397

Huerta-Cepas, J., S. Capella-Gutiérrez, L.P. Pryszcz, M. Marcet-Houben, and T. Gabaldón. 2014. PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res. 42:D897–D902. doi:10.1093/nar/gkt1177

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

Jaillon, O., J.M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467. doi:10.1038/nature06148

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

Johnston, J.S., A.E. Pepper, A.E. Hall, Z.J. Chen, G. Hodnett, J. Drabek, R. Lopez, and H.J. Price. 2005. Evolution of genome size in Brassicaceae. Ann. Bot. (Lond.) 95:229–235. doi:10.1093/aob/mci016

Kersey, P.J., J.E. Allen, I. Armean, S. Boddu, B.J. Bolt, D. Carvalho-Silva, et al. 2015. Ensembl genomes 2016: More genomes, more complexity. Nucleic Acids Res. 44:D574–D580. doi:10.1093/nar/gkv1209

Lee, T.H., H. Tang, X. Wang, and A.H. Paterson. 2013. PGDD: A database of gene and genome duplication in plants. Nucleic Acids Res. 41:D1152–1158. doi:10.1093/nar/gks1104

Letunic, I., and P. Bork. 2011. Interactive tree of life v2: Online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 39:W475–W458. doi:10.1093/nar/gkr201

Li, L., G. Ji, C. Ye, C. Shu, J. Zhang, and C. Liang. 2015. PlantOrDB: A genome-wide ortholog database for land plants and green algae. BMC Plant Biol. 15:161. doi:10.1186/s12870-015-0531-4

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

Lynn, D.J., G.L. Winsor, C. Chan, N. Richard, M.R. Laird, A. Barsky, et al. 2008. InnateDB: Facilitating systems-level analyses of the mammalian innate immune response. Mol. Syst. Biol. 4:218. doi:10.1038/msb.2008.55

Monaco, M.K., J. Stein, S. Naithani, S. Wei, P. Dharmawardhana, S. Kumari, V. Amarasinghe, et al. 2014. Gramene 2013: Comparative plant genomics resources. Nucleic Acids Res. 42:D1193–D1199. doi:10.1093/nar/gkt1110

Ostlund, G., T. Schmitt, K. Forslund, T. Köstler, D.N. Messina, S. Roopra, O. Frings, and E.L.L. Sonnhammer. 2010. InParanoid 7: New algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 38:D196–D203. doi:10.1093/nar/gkp931

Ozdemir, B.S., P. Hernandez, E. Filiz, and H. Budak. 2008. *Brachypodium* genomics. Int. J. Plant Genomics 2008:536104. doi:10.1155/2008/536104

Palenik, B., J. Grimwood, A. Aerts, P. Rouzé, A. Salamov, N. Putnam, et al. 2007. The tiny eukaryote ostreococcus provides genomic insights into the paradox of plankton speciation. Proc. Natl. Acad. Sci. U.S.A. 104:7705–7710. doi:10.1073/pnas.0611046104

Paterson, A.H., J.E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, G. Haberer, et al. 2009. The *Sorghum bicolor* genome and the diversification of grasses." Nature 457:551–556. doi:10.1038/nature07723

Powell, S., K. Forslund, D. Szklarczyk, K. Trachana, A. Roth, J. Huerta-Cepas, et al. 2014. eggNOG v4.0: Nested orthology inference across 3686 organisms. Nucleic Acids Res. 42:D231–D239. doi:10.1093/nar/gkt1253

Proost, S., M. Van Bel, D. Vaneechoutte, Y. Van de Peer, D. Inzé, B. Mueller-Roeber, and K. Vandepoele. 2015. PLAZA 3.0: An access point for plant comparative genomics. Nucleic Acids Res. 43:D974–D981. doi:10.1093/nar/gku986

Pryszcz, L.P., J. Huerta-Cepas, and T. Gabaldón. 2011. MetaPhOrs: Orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res. 39:E32. doi:10.1093/nar/gkq953

Remm, M., C.E. Storm, and E.L. Sonnhammer. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J. Mol. Biol. 314:1041–1052. doi:10.1006/jmbi.2000.5197

Ruan, J., H. Li, Z. Chen, A. Coghlan, L.J.M. Coin, Y. Guo, et al. 2008. TreeFam: 2008 update. Nucleic Acids Res. 36:D735–D740. doi:10.1093/nar/gkm1005

Schmutz, J., S.B. Cannon, J. Schlueter, J. Ma, T. Mitros, W. Nelson, et al. 2010. Genome sequence of the palaeopolyploid soybean. Nature 463:178–183. doi:10.1038/nature08670

Schnable, P.S., D. Ware, R.S. Fulton, J.C. Stein, F. Wei, S. Pasternak, et al. 2009. The B73 maize genome: Complexity, diversity, and dynamics. Science 326:1112–1115. doi:10.1126/science.1178534

Silvar, C., M.M. Martis, T. Nussbaumer, N. Haag, R. Rauser, J. Keilwagen, V. Korzun, K.F.X. Mayer, F. Ordon, and D. Perovic. 2015. Assessing the barley genome zipper and genomic resources for breeding purposes. Plant Genome 8. doi:10.3835/plantgenome2015.06.0045

Tang, H., V. Krishnakumar, S. Bidwell, B. Rosen, A. Chan, S. Zhou, L. Gentzbittel, et al. 2014. An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. BMC Genomics 15:312. doi:10.1186/1471-2164-15-312

Tatusov, R.L., E.V. Koonin, and D.J. Lipman. 1997. A genomic perspective on protein families. Science 278:631–637. doi:10.1126/science.278.5338.631

Tatusov, R.L., N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, et al. 2003. The COG database: An updated version includes eukaryotes. BMC Bioinf. 4:41. doi:10.1186/1471-2105-4-41

Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641. doi:10.1038/nature11119

Tulpan, D., S. Leger, A. Tchagang, and Y. Pan. 2015. Enrichment of *Triticum aestivum* gene annotations using ortholog cliques and gene ontologies in other plants. BMC Genomics 16:299. doi:10.1186/s12864-015-1496-2

van der Heijden, R.T.J.M., B. Snel, V. van Noort, and M.A. Huynen. 2007. Orthology prediction at scalable resolution by phylogenetic tree analysis. BMC Bioinf. 8:83. doi:10.1186/1471-2105-8-83

Wang, Y., D. Coleman-Derr, G. Chen, and Y.Q. Gu. 2015. OrthoVenn: A web server for genome wide comparison and annotation of orthologous clusters across multiple species. Nucleic Acids Res. 43:W78–W84. doi:10.1093/nar/gkv487

Waterhouse, R.M., F. Tegenfeldt, J. Li, E.M. Zdobnov, and E.V. Kriventseva. 2013. OrthoDB: A hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Res. 41:D358–D365. doi:10.1093/nar/gks1116

Wicker, T., W. Zimmermann, D. Perovic, A.H. Paterson, M. Ganal, A. Graner, and N. Stein. 2004. A detailed look at 7 million years of genome evolution in a 439 kb contiguous sequence at the barley Hv-eIF4E locus: Recombination, rearrangements and repeats." Plant J. 41:184–194. doi:10.1111/j.1365-313X.2004.02285.x

Wolf, Y.I., and E.V. Koonin. 2012. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. Genome Biol. Evol. 4:1286–1294. doi:10.1093/gbe/evs100

Xu, X., S. Pan, S. Cheng, B. Zhang, D. Mu, P. Ni, et al. 2011. Genome sequence and analysis of the tuber crop potato. Nature 475:189–195. doi:10.1038/nature10158

Zhang, G., X. Liu, Z. Quan, S. Cheng, X. Xu, S. Pan, .et al. 2012. Genome sequence of foxtail millet (*Setaria italica*) provides insights into grass evolution and biofuel potential. Nature Biotech. 30:549–554. doi:10.1038/nbt.2195