The Plant Cell, Vol. 21: 3718–3731, December 2009, www.plantcell.org

PERSPECTIVE
Special Series on Large-Scale Biology

PLAZA: A Comparative Genomics Resource to Study Gene and Genome Evolution in Plants [W]

Sebastian Proost,[a,b,1] Michiel Van Bel,[a,b,1] Lieven Sterck,[a,b] Kenny Billiau,[a,b] Thomas Van Parys,[a,b] Yves Van de Peer,[a,b,2] and Klaas Vandepoele[a,b]

[a] Department of Plant Systems Biology, Flanders Institute for Biotechnology, B-9052 Ghent, Belgium
[b] Department of Molecular Genetics, Ghent University, B-9052 Ghent, Belgium

The number of sequenced genomes of representatives within the green lineage is rapidly increasing. Consequently, comparative sequence analysis has significantly altered our view on the complexity of genome organization, gene function, and regulatory pathways. To explore all this genome information, a centralized infrastructure is required where all data generated by different sequencing initiatives is integrated and combined with advanced methods for data mining. Here, we describe PLAZA, an online platform for plant comparative genomics (http://bioinformatics.psb.ugent.be/plaza/). This resource integrates structural and functional annotation of published plant genomes together with a large set of interactive tools to study gene function and gene and genome evolution. Precomputed data sets cover homologous gene families, multiple sequence alignments, phylogenetic trees, intraspecies whole-genome dot plots, and genomic colinearity between species. Through the integration of high confidence Gene Ontology annotations and tree-based orthology between related species, thousands of genes lacking any functional description are functionally annotated. Advanced query systems, as well as multiple interactive visualization tools, are available through a user-friendly and intuitive Web interface. In addition, detailed documentation and tutorials introduce the different tools, while the workbench provides an efficient means to analyze user-defined gene sets through PLAZA's interface. In conclusion, PLAZA provides a comprehensible and up-to-date research environment to aid researchers in the exploration of genome information within the green plant lineage.

INTRODUCTION

The availability of complete genome sequences has significantly altered our view on the complexity of genome organization, genome evolution, gene function, and regulation in plants. Whereas large-scale cDNA sequencing projects have generated detailed information about gene catalogs expressed in different tissues or during specific developmental stages (Rudd, 2003), the application of genome sequencing combined with high-throughput expression profiling has revealed the existence of thousands of unknown expressed genes conserved within the green plant lineage (Gutierrez et al., 2004; Vandepoele and Van de Peer, 2005). The generation of high-quality complete genome sequences for the model species *Arabidopsis thaliana* and rice (*Oryza sativa*) required large international consortia and took several years before completion (Arabidopsis Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005). Facilitated by whole-genome shotgun and next-generation sequencing technologies, genome information for multiple plant species is now rapidly expanding. The genomes of four eudicots, *Arabidopsis*, poplar (*Populus trichocarpa*), grapevine (*Vitis vinifera*), and papaya (*Carica papaya*), two monocots, rice and *Sorghum bicolor*, the moss *Physcomitrella patens*, and several green algae (Parker et al., 2008) have been published, and new genome initiatives will at least double the number of plant genome sequences by the end of this decade (Paterson, 2006; Pennisi, 2007).

Although the genomes of some of these species provide invaluable resources as economical model systems, comparative analysis makes it possible to learn more about the different characteristics of each organism and to link phenotypic with genotypic properties. Hanada and coworkers demonstrated how the integration of expression data and multiple plant sequences combined with evolutionary conservation can greatly improve gene discovery (Hanada et al., 2007; Brady and Provart, 2009). Whereas a detailed gene catalog provides a starting point to study growth and development in model organisms, sequencing species from different taxonomic clades generates an evolutionary framework to study how changes in coding and

[1] These authors contributed equally to this work.
[2] Address correspondence to yves.vandepeer@psb.ugent.be.
The authors responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) are: Sebastian Proost (sebastian.proost@psb.vib-ugent.be) and Klaas Vandepoele (klaas.vandepoele@psb.vib-ugent.be).
[W] Online version contains Web-only data.
www.plantcell.org/cgi/doi/10.1105/tpc.109.071506

PERSPECTIVE

noncoding DNA affect the evolution of genes, resulting in expression divergence and species-specific adaptations (Tanay et al., 2005; Blomme et al., 2006; Stark et al., 2007). Based on orthologous genes (i.e., genes sharing common ancestry evolved through speciation), comparative genomics provides a powerful approach to exploit mapping data, sequence information, and functional information across various species (Fulton et al., 2002). Similarly, the analysis of genes or pathways in a phylogenetic context allows scientists to better understand how complex biological processes are regulated and how morphological innovations evolve at the molecular level. For example, studying gene duplicates in poplar has revealed specific expansions in gene families related to cell wall formation covering cellulose and lignin biosynthesis genes and genes associated with disease and insect resistance (Tuskan et al., 2006). Similarly, amplifications of genes belonging to the metabolic pathways of terpenes and tannins in grapevine directly relate the diversity of wine flavors with gene content (Jaillon et al., 2007). Besides the comparative analysis of specific gene families in higher plants, comparisons with other members of the green lineage provide additional information about the evolutionary processes that have changed gene content during hundreds of millions of years. Although the genomes of, for instance, moss and green algae contain a smaller number of genes compared with flowering plants, they provide an excellent starting point to reconstruct the ancestral set of genes at different time points during plant evolution and to trace back the origin of newly acquired genes (Merchant et al., 2007; Rensing et al., 2008).

Gene duplication has been extensive in plant genomes. In addition, detailed comparison of gene organization and genome structure has identified multiple whole-genome duplication (WGD) events in different land plants. From a biological point of view, the large number of small- and large-scale duplication events in flowering plants has had a great influence on the evolution of gene function and regulation. For instance, between 64 and 79% of all protein-coding genes in *Arabidopsis*, poplar, and rice are part of multigene families, compared with 40% for the green alga *Chlamydomonas reinhardtii*. Paralogs are generally considered to evolve through nonfunctionalization (silencing of one copy), neofunctionalization (acquisition of a novel function for one copy), or subfunctionalization (partitioning of tissue-specific patterns of expression of the ancestral gene between the two copies) (Conant and Wolfe, 2008; Freeling, 2009). The impact of the large number of duplicates on the complexity, redundancy, and evolution of regulatory networks in multicellular organisms is currently far from being well understood (Chen, 2007; Rosin and Kramer, 2009).

Performing evolutionary and comparative analyses to study gene families and genome organization requires a centralized plant genomics infrastructure where all information generated by different sequencing initiatives is integrated, in combination with advanced methods for data mining. Even though general formats have been developed to store and exchange gene annotation (Stein, 2001), the properties of available plant genomic data (i.e., structural annotation of protein-coding genes, RNAs, transposable elements, pseudogenes, or functional annotations through protein domains or ontologies) vary greatly between different sequencing centers, impeding comparative analyses for nonexpert users. Additionally, large-scale comparisons between multiple eukaryotic species require huge computational resources to process the large amounts of data. Here, we present PLAZA, a new online resource for plant comparative genomics (http://bioinformatics.psb.ugent.be/plaza/). We show how PLAZA provides a versatile platform for integrating published plant genomes to study gene function and genome evolution. Precomputed comparative genomics data sets cover homologous gene families, multiple sequence alignments, phylogenetic trees, intraspecies whole-genome dot plots, and genomic colinearity information between species. Multiple visualization tools that are available through a user-friendly Web interface make PLAZA an excellent starting point to translate sequence information into biological knowledge.

Data Assembly

The current version of PLAZA contains the nuclear and organelle genomes of nine species within the *Viridiplantae* kingdom: the four eudicots *Arabidopsis*, papaya, poplar, and grapevine, the two monocots rice and sorghum, the moss *P. patens*, and the unicellular green algae *C. reinhardtii* and *Ostreococcus lucimarinus*. The integration of all gene annotations provided by the different sequencing centers yielded a data set of 295,865 gene models, of which 92.6% represent protein-coding genes (Table 1). The remaining genes are classified as transposable elements, RNA, and pseudogenes (6.5, 0.6, and 0.3%, respectively). Whereas most of the genes are encoded in the nuclear genomes, a small set are from chloroplast and mitochondrial origin (0.4 and 0.2%, respectively). For all genes showing alternative splicing, the longest transcript was selected as a reference for all downstream comparative genomics analyses. Detailed gene annotation, including information about alternative splicing variants is displayed using the AnnoJ genome browser (Lister et al., 2008). Whereas genomes from model species like *Arabidopsis* and rice are characterized by high sequence coverage and a set of contiguous genomic sequences resembling the actual number of chromosomes, other genome sequences, such as those of *P. patens* and papaya, are produced by the whole-genome shotgun sequencing method and contain more than 1000 genomic scaffolds (Table 1). For poplar, grape, and sorghum, a large fraction of the genome is assembled into chromosomes, but several scaffolds that could not be anchored physically are still present in the data set. In this case, we allocated the genes that were not assigned to a chromosome in the original annotation to a virtual chromosome zero. This procedure reduces the number of pseudomolecules when applying genome evolution studies while preserving the correct proteome size (i.e., the total number of proteins per species) and the relative gene positions on the genomic scaffolds (Table 1).

PERSPECTIVE

Table 1. Summary of the Gene Content in PLAZA

Species	Genome Size[a]	Genes[b]		Scaffolds[c]	Coding	GO[d]	InterPro
Arabidopsis thaliana	115 Mb (BAC/PAC/TAC)	33,284	81.81%	5	27,228	63.62% (66.21%)	56.49%
Carica papaya	271 Mb (3× WGS)	28,072	99.84%	1,898	28,027	0.00% (22.88%)	57.75%
Populus trichocarpa	410 Mb (7.5× WGS)	45,699	99.90%	19+1 (5724)	45,654	44.69% (52.89%)	61.91%
Vitis vinifera	468 Mb (8.4× WGS)	38,127	99.63%	19+1 (35)	37,987	40.09% (45.90%)	57.62%
Oryza sativa	371 Mb (BAC/PAC)	57,955	72.32%	12	41,912	30.42% (30.91%)	63.69%
Sorghum bicolor	626 Mb (WGS)	34,686	99.78%	10+1 (217)	34,609	44.44% (48.13%)	67.79%
Physcomitrella patens	480 Mb (8.6× WGS)	36,137	99.80%	1,446	36,065	33.20%	42.44%
Chlamydomonas reinhardtii	121 Mb (13× WGS)	14,731	99.64%	552	14,678	34.99%	49.29%
Ostreococcus lucimarinus	13 Mb (WGS)	7,805	100.00%	21	7,805	47.94%	62.86%
Total		295,865	92.60%		273,965	39.36%	44.88%

[a]Size assembled (sequencing method). PAC, phague artificial chromosome; TAC, transformation-competent artificial chromosome; WGS, whole-genome shotgun.
[b]Percentage of protein-coding genes.
[c]Numbers in parentheses refer to the number of genomic sequences in the original annotation; "+1" indicates the creation of a virtual chromosome zero to group scaffolds
[d]Percentages in parentheses include projected GO annotations, while the first value only reports original primary GO data.

Complementary to the structural annotation, we also retrieved, apart from free-text gene descriptions, functional information through Gene Ontology (GO) associations (Ashburner et al., 2000), InterPro domain annotations (Hunter et al., 2009), and *Arabidopsis* Reactome pathway data (Tsesmetzis et al., 2008). Whereas GO provides a controlled vocabulary to describe gene and gene product attributes (using Cellular Component, Biological Process, and Molecular Function), the InterPro database provides an annotation system in which identifiable features found in known proteins (i.e., protein families, domains, and functional sites) can be applied to new protein sequences. GO provides a set of different evidence codes that indicate the nature of the evidence that supports a particular annotation. The *Arabidopsis* Reactome is a curated resource for pathways where enzymatic reactions are added to genes and a set of reactions is grouped into a pathway.

Apart from the basic information related to gene structure and function (e.g., genome coordinates, mRNA coding and protein sequences, protein domains, and gene description), different types of comparative genomics information are provided through a variety of Web tools. In general, these data and methods can be classified as approaches to study gene homology and genome structure within and between species. Whereas the former focuses on the organization and evolution of families covering homologous genes, the latter exploits gene colinearity, or the conservation of gene content and order, to study the evolution of plant genomes (Figure 1).

Delineating Gene Families and Subfamilies

As a starting point to study gene function and evolution, all protein-coding genes are stored in gene families based on sequence similarity inferred through BLAST (Altschul et al., 1997). A gene family is defined as a group of two or more homologous genes. A graph-based clustering method (Markov clustering implemented in Tribe-MCL; Enright et al., 2002) was used to delineate gene families based on BLAST protein similarities in a process that is sensitive to the density and the strength of the BLAST hits between proteins. Although this method is very well suited for clustering large sets of proteins derived from multiple species, high false-positive rates caused by the potential inclusion of spurious BLAST hits have been reported (Chen et al., 2007). Therefore, we applied a postprocessing procedure by tagging genes as outliers if they showed sequence similarity to only a minority of all family members (see Supplemental Methods 1 online). The OrthoMCL method (Li et al., 2003) was applied to build subfamilies based on the same protein similarity graph. Benchmark experiments have shown that OrthoMCL yields fewer false positives compared with the Tribe-MCL method and that, overall, it generates tighter clusters containing a smaller number of genes (Chen et al., 2007). Because OrthoMCL models orthology and in-paralogy (duplication events after dating speciation) based on a reciprocal-best hit strategy, the final protein clusters will be smaller than Tribe-MCL clusters because out-paralogs (homologs from duplication events predating speciation) will not be grouped. Therefore, from a biological point of view, subfamilies or out-paralogs can be considered as different subtypes within a large protein family.

In total, 77.62% of all protein-coding genes (212,653 genes) are grouped in 14,742 multigene families, leaving 61,312 singleton genes (see Supplemental Table 1 online). Sixty-two percent of these families cover genes from multiple species, and for approximately one-fifth, multiple subfamilies were identified. Manual inspection and phylogenetic analysis of multiple families revealed that in many cases, OrthoMCL correctly identified out-paralogous groups that can be linked with distinct biological subtypes or functions (see Supplemental Methods 2 online; Hanada et al., 2008). Examples of identified subfamilies are

PERSPECTIVE

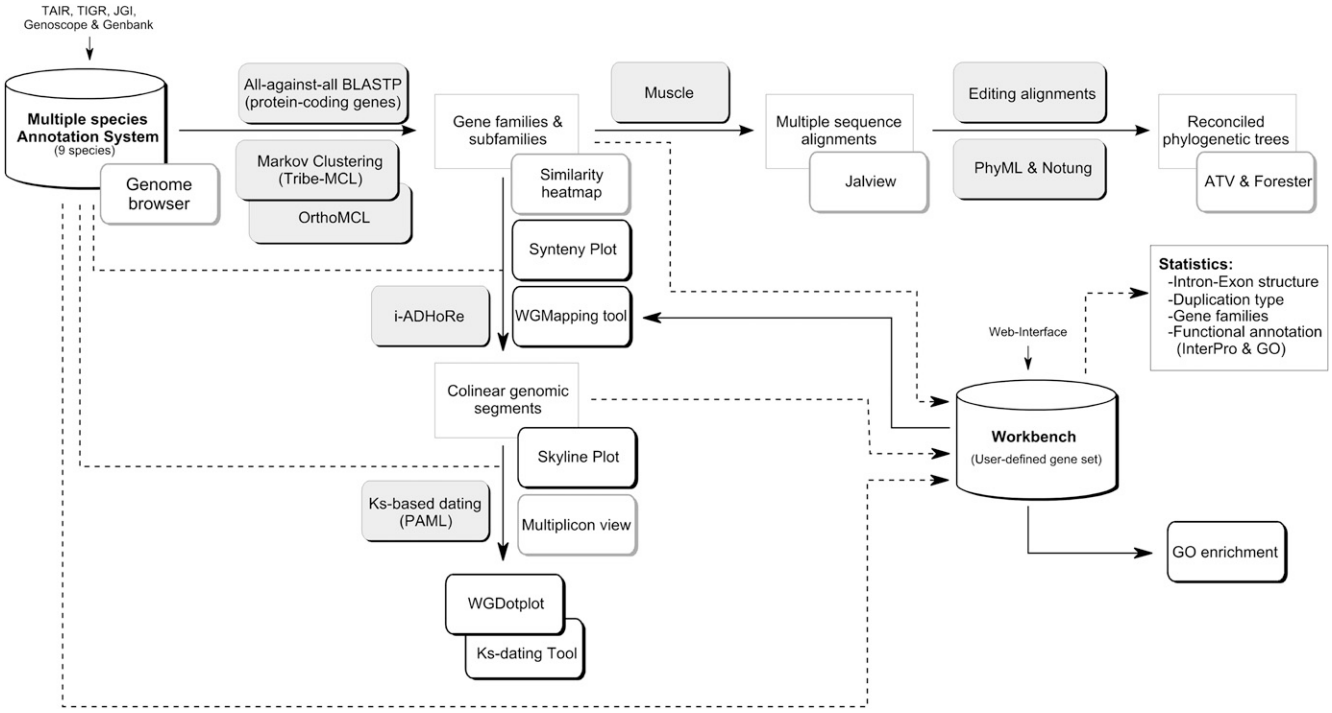

Figure 1. Structure of the PLAZA Platform.

Outline of the different data types (white boxes) and tools (gray rounded boxes) integrated in the PLAZA platform. White rounded boxes indicate the different tools implemented to explore the different types of data available through the website.

different clathrin adaptors (Adaptor Protein complex subunits), minichromosome maintenance subunits, ATP binding GCN transporters, cullin components of SCF ubiquitin ligase complexes, replication factors, and $\alpha/\beta/\gamma$ tubulins (Figure 2; see Supplemental Table 2 online). Although fast-evolving genes or homologs showing only limited sequence similarity can lead to incorrect families, a similarity heat map tool was developed to explore all pairwise sequence similarities per family (Figure 2). This visualization provides an intuitive approach, complementary to the automatic protein clustering and phylogenetic trees, to explore gene homology. In addition, a BLAST interface is available that provides a flexible entry point to search for homologous genes using user-defined sequences and parameter settings.

Phylogenetic Inference and the Projection of Functional Annotation via Orthology

Phylogenetic studies generate valuable information on the evolutionary and functional relationships between genes of different species, genomic complexity, and lineage-specific adaptations. In addition, they provide an excellent basis to infer orthology and paralogy (Koonin, 2005). Based on the gene families generated using protein clustering, a phylogenetic pipeline was applied to construct 20,781 phylogenetic trees covering ~172,000 protein-coding genes (see Supplemental Table 1 online). Bootstrapped phylogenetic trees were constructed using the maximum likelihood method PhyML (Guindon and Gascuel, 2003) based on protein multiple sequence alignments generated using MUSCLE (Edgar, 2004) (see Supplemental Methods 3 online). In order to extract biological information from all phylogenies, we applied the NOTUNG tree reconciliation method to annotate, based on parsimony and a species tree, tree nodes as duplication/speciation events together with a time estimate (Vernot et al., 2008). Detailed inspection of tree topologies revealed that, even for well-supported nodes with high bootstrap values, a high number of nodes (53 to 64%) correspond with falsely inferred duplication events (see Supplemental Figure 1 online). This problem is caused by the different rates of amino acid evolution in different species, potentially leading to incorrect evolutionary reconstructions (Hahn, 2007). Therefore, we calculated a duplication consistency score, originally developed by Ensembl (Vilella et al., 2009), to identify erroneously inferred duplication events (see Supplemental Methods 3 and Supplemental Figure 1 online). This score reports, for a duplication node, the intersection of the number of postduplication species over the union and is typically high for tree nodes denoting a real duplication event. Consequently, the reconciled phylogenetic trees provide a reliable means to identify biologically relevant duplication and

PERSPECTIVE

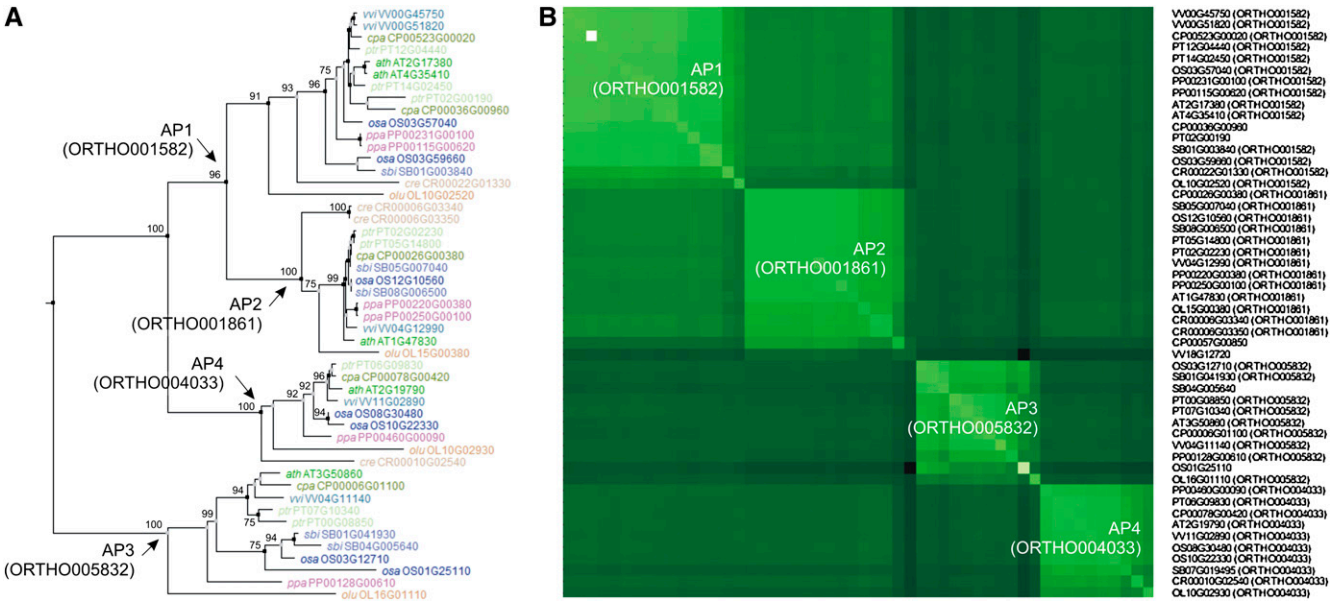

Figure 2. Gene Family Delineation Using Protein Clustering, Phylogenetic Tree Construction, and Similarity Heat Maps.

(A) Phylogenetic tree of clathrin adaptors (HOM000575) with the AP1-4 subfamilies delineated using OrthoMCL. Black and gray squares on the tree nodes indicate duplication and speciation events identified using tree reconciliation, respectively. Only bootstrap values ≥70% are shown.
(B) Similarity heat map displaying all pairwise similarity scores for all gene family members. BLAST bit scores were converted to a color gradient with white/bright green and dark green indicating high and low scores, respectively. Clustering of the sequence similarities supports the existence of the four AP subfamilies that were identified using protein clustering and confirmed using phylogenetic inference. Note that subfamilies AP3 and AP4 are inverted in the heat map compared with the tree. Species abbreviations as in Table 2.

speciation events (or paralogs and orthologs, respectively). In addition, the time estimates at each node make it possible to infer the age of paralogs and correlate duplications with evolutionary adaptations.

Since speciation events inferred through phylogenetic tree construction provide a reliable way to identify orthologous genes, these orthology relationships can be used to transfer functional annotation between related organisms (Hubbard et al., 2005; Tsesmetzis et al., 2008; The Reference Genome Group of the Gene Ontology Consortium, 2009). We applied a stringent set of rules to identify a set of eudicot and monocot tree-based orthologous groups and used GO projection to exchange functional annotation between species (see Supplemental Methods 4 and Supplemental Figure 2 online). Whereas in the original annotation, 39% of all proteins were annotated with at least one GO term, this fraction greatly varies for different species (Table 1). Model species like *Arabidopsis* and rice have a large set of functionally annotated genes with GO terms supported by various experimentally derived evidence codes. By contrast, other organisms only have annotations inferred through electronic annotation (e.g., grapevine and popular) or completely lack functional annotation (e.g., papaya; see Supplemental Table 3 online). Application of GO projection using eudicot and monocot orthologous groups resulted in new or improved functional information for 36,473 genes. This projected information covers ~105,000 new annotations, of which one-fifth is supported by evidence from multiple genes. Overall, 11.8% of all genes lacking GO information in flowering plants could be annotated based on functional data of related genes/species and for ~22,000 genes (17% of protein-coding genes in angiosperms already annotated using GO) new or more specific GO terms could be assigned. For papaya, initially lacking functional GO data, 39% of all genes for which a phylogenetic tree exists have now one or more associated GO term (see Supplemental Table 3 online). To estimate the specificity of the functional annotations, we used the GO depth (i.e., the number of shortest-path-to-root steps in the GO hierarchy) as a measure for the information content for the different annotations. Distributions per species reveal that the projected annotations are as detailed as the original primary GO data and that for species initially lacking GO information, detailed GO terms can be associated to most genes (see Supplemental Table 4 online). Whereas Blast2GO, a high-throughput and automatic functional annotation tool (Gotz et al., 2008), applies sequence similarity to identify homologous genes and collect primary GO data, GO projection uses phylogenetic inference to identify orthologous genes prior to transfer of functional annotation. Both methods incorporate information from different GO evidence tags to avoid the inclusion of low-quality annotations while

generating functional information for uncharacterized proteins. It is important to note that all pages and tools presenting functional annotation through the PLAZA website can be used, including either all GO data or only the primary GO annotations (i.e., excluding projected GO terms).

Exploring Genome Evolution in Plants

To study plant genome evolution, PLAZA provides various tools to browse genomic homology data, ranging from local synteny to gene-based colinearity views. Whereas colinearity refers to the conservation of gene content and order, synteny is more loosely defined as the conservation of similar genes over two or more genomic regions. Moreover, genome organization can be explored at different levels, making it possible to easily navigate from chromosome-based views to detailed gene-centric information for one or multiple species. Based on gene family delineation and the conservation of gene order, homologous genomic regions were detected using i-ADHoRe (Simillion et al., 2008). The i-ADHoRe algorithm combines gene content and gene order information within a statistical framework to find significant microcolinearity taking into account different types of local rearrangements (Vandepoele et al., 2002). Subsequently, these colinear regions are used to build genomic profiles that allow the identification of additional homologous segments. As a result, sets of homologous genomic segments are grouped into what is referred to as a multiplicon. The multiplication level indicates the number of homologous segments for a given genomic region. The advantage of profile searches (also known as top-down approaches) is that degenerate colinearity (or ancient duplications) can still be detected (Vandepoele et al., 2002; Simillion et al., 2004).

The Synteny plot is the most basic tool to study gene-centric genomic homology. This feature shows all genes from the specified gene family with their surrounding genes, providing a less stringent criterion to study genomic homology compared with colinearity. To ensure the fast exploration of positional orthologs, gene family members have been clustered based on their flanking gene content (see Supplemental Figure 3 online). Investigating colinearity on a genome-wide scale can be done using the WGDotplot (Figure 3A). This tool can be applied to identify large-scale duplications within a genome or to study genomic rearrangements within or between species (e.g., after genome doubling or speciation, respectively). In a first view, a genome-wide plot displays inter- or intraspecies colinearity, while various features are available to zoom in to chromosome-wide plots or the underlying multiplicon gene order alignment. Intraspecies comparisons can also be visualized using circular plots that depict all duplicated blocks physically mapped on the chromosomes.

All colinear gene pairs (or block duplicates) have been dated using K_s, the synonymous substitution rate (see Supplemental Methods 6 online). K_s is considered to evolve at a nearly constant neutral rate since synonymous substitutions do not alter the encoded amino acid sequence. As a consequence, these values can be used as a molecular clock for dating, although saturation (i.e., when synonymous sites have been substituted multiple times, resulting in K_s values >1) can lead to underestimation of the actual age (Smith and Smith, 1996). The average K_s for a colinear (or duplicated) block is calculated and colored accordingly in the WGDotplots (Figure 3A). Based on the K_s distributions of block paralogs, the K_s dating tool can be employed to date one or more large-scale duplication events relative to a speciation event considering multiple species. As shown in Supplemental Figure 4 online, ancient and more recent WGDs can be identified in several plants species, although varying evolutionary rates in different lineages due to, for instance, different generation times, might interfere with the accurate dating of these events (Tang et al., 2008a; Van de Peer et al., 2009).

When investigating genomic homology between more than two genomes, the Skyline plot provides a rapid and flexible way to browse multiple homologous genomic segments (Figure 3B). For a region centered around a reference gene, all colinear segments (from the selected set of organisms) are determined and visualized using color-coded stacked segments. The Skyline plot offers a comprehensive view of the number of regions that are colinear in the species selected (see Supplemental Methods 5 online). Navigation buttons allow the user to scroll left and right, whereas a window size parameter setting provides a zooming function to focus either on a small region around the reference gene or on the full chromosome. Clicking on one of the regions of interest shows a more detailed view (Multiplicon view; see Figure 3C). The gene alignment algorithm maintains the original gene order but will introduce gaps to place homologous genes in the same column (if possible).

Database Access, User Interface, and Documentation

An advanced query system has been developed to access the different data types and research tools and to quickly retrieve relevant information. Starting from a keyword search on gene descriptions, GO terms, InterPro domains, Reactome pathways, or a gene identifier, relevant genes and gene families can be fetched. Apart from the internal PLAZA gene identifiers, the original gene names provided by the data provider are supported as well. When multiple genes are returned using the search function, the "view-associated gene families" option makes it possible to link all matching genes to their corresponding gene families, reducing the complexity of the number of returned items. When searching for genes related to a specific biological process using GO, this function makes it possible to directly identify all relevant gene families and analyze the evolution of these genes in the different species. Although for some species the functional annotation is limited, even after GO projection, mapping genes related to a specific functional category to the corresponding families makes it possible to rapidly explore functional annotations in different species through gene homology.

PERSPECTIVE

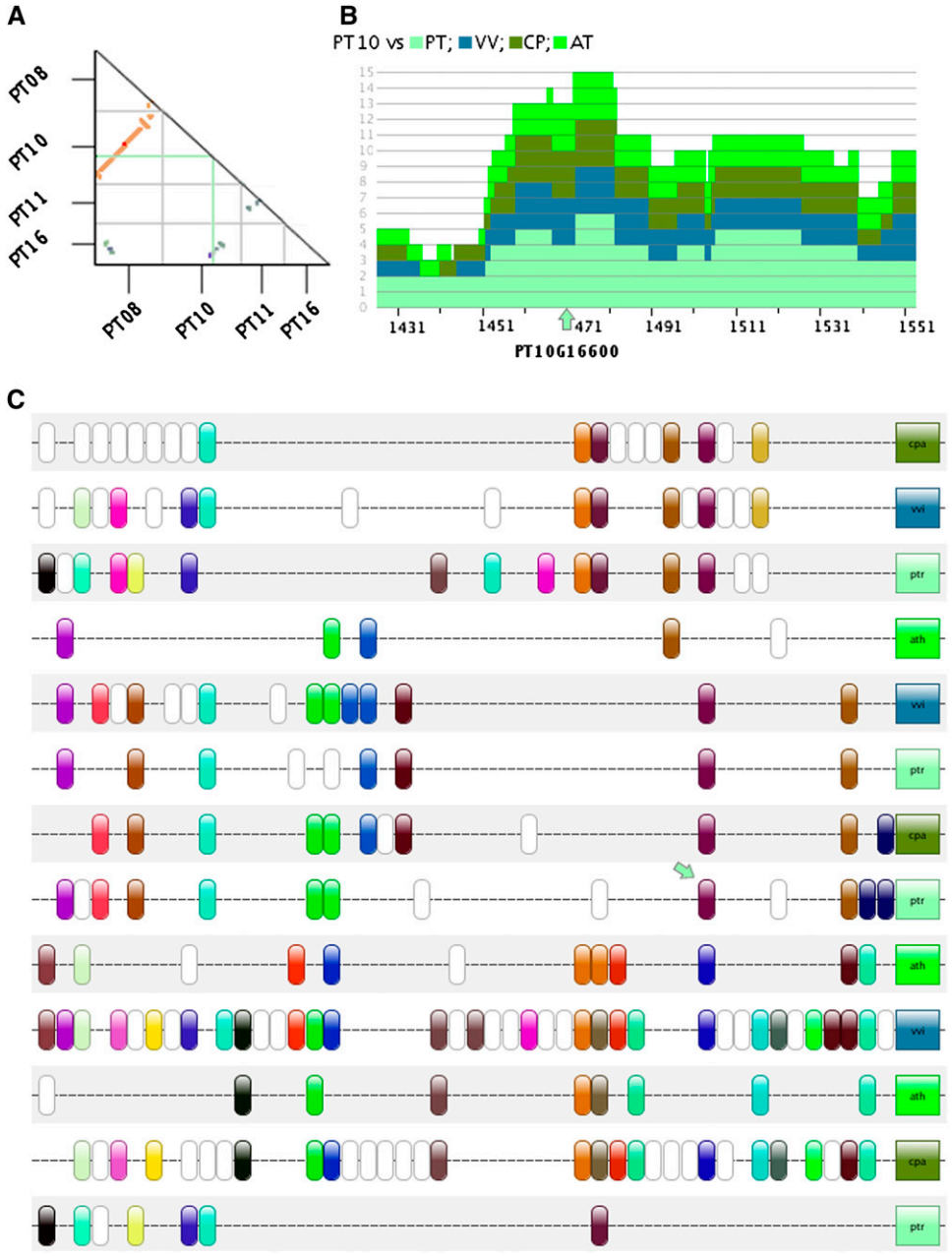

Figure 3. Overview of Different Colinearity-Based Visualizations of the Genomic Region around Poplar Gene PT10G16600.

(A) The WGDotplot shows that the gene of interest, indicated by the light-green line, is located in a duplicated block between chromosomes PT08 and PT10. The orange color refers to a K_s value of 0.2 to 0.3, indicating the most recent WGD in poplar.
(B) The Skyline plot shows the number of colinear segments in different organisms detected using i-ADHoRe.
(C) The Multiplicon view depicts the gene order alignment of the homologous segments indicated in **(B)**. Whereas the rounded boxes represent the different genes color-coded according to the gene family they belong to, the square boxes at the right indicate the species the genomic segment was sampled from. The reference gene is indicated by the light-green arrow in **(B)** and **(C)**.

PERSPECTIVE

To analyze multiple genes in batch, we have developed a Workbench where, for user-defined gene sets, different genome statistics can be calculated (Figure 1). Genes can be uploaded through a list of (internal or external) gene identifiers or based on a sequence similarity search. For example, this last option enables users to map an EST data set from a nonmodel organism to a reference genome annotation present in PLAZA. For gene sets saved by the user in the Workbench detailed information about functional annotation (InterPro and GO), associated gene families, block and tandem gene duplicates, and gene structure are provided. In addition, the GO enrichment tool allows for determination of whether a user-defined gene set is overrepresented for one or more GO terms (see the Workbench tutorial on the PLAZA documentation page). This feature makes it possible to rapidly explore functional biases present in, for example, differentially expressed genes or EST libraries.

The organization of a gene set of interest (e.g., gene family homologs, genes with a specific InterPro domain, GO term, or from a Reactome pathway, a Workbench gene set) in a genome-wide context can reveal interesting information about genomic clustering. The Whole Genome Mapping tool can be used to display a selection of genes on the chromosomes (see Supplemental Figure 5 online), and additional information about the duplication type of these genes (i.e., tandem or block duplicate) is provided. Furthermore, the Whole Genome Mapping tool allows users to view the distribution of different gene types (protein-coding, RNA, pseudogene, or transposable element) per species.

An extensive set of documentation pages describes the sources of all primary gene annotations, the different methods and parameters used to build all comparative genomics data, and instructions on how to use the different tools. We also provide a set of tutorials introducing the different data types and interactive research tools. An extensive glossary has been compiled that interactively is shown on all pages when hovering over specific terms. Finally, for each data type (e.g., gene family and GO term) or analysis tool, all data can be downloaded as simple tab-delimited text files. Bulk downloads covering sequence or annotation data from one or more species are available through an FTP server.

Data Analysis: Dissecting Plant Gene Duplicates Using PLAZA

To illustrate the applicability of PLAZA for comparative genomics studies, a combination of tools was used to characterize in detail the mode and tempo of gene duplications in plants. In the first case study, tree-based dating and GO enrichment analysis were used to analyze the gene functions of species-specific paralogs. Initially, gene duplicates were extracted from the reconciled phylogenetic trees for all organisms. To ensure the reliability of the selected duplication nodes, we only retained nodes with good bootstrap support ($\geq$70%) and consistency scores (>0). By cross-referencing all returned genes with the colinearity information included in PLAZA, all species-specific duplicates were further divided into tandem and block duplicates. Subsequently, enriched GO terms were calculated for each of those gene sets using PLAZA's workbench.

Whereas in the green alga *O. lucimarinus*, 45% of all species-specific duplicates are derived from a recent segmental duplication between chromosomes 13 and 21, nearly half of all *Vitis*-specific duplicates correspond with tandem duplications (see Supplemental Table 5 online). For many species, tandem duplications account for the largest fraction (34 to 50%) of species-specific paralogs. The GO enrichment analysis provides an efficient approach to directly relate duplication modes in different species with specific biological processes or evolutionary adaptations. Browsing the associated gene families makes it possible to explore the functions of the different genes (Figure 4). For example, the GO term "response to biotic stimulus" (GO:0009607) was enriched for the tandem duplicates of *Arabidopsis*, poplar, and grapevine. When focusing on the duplicated genes causing this enrichment, we observed that different gene families involved in biotic response are expanded in different species (Figure 4B). Whereas in *Arabidopsis*, the Avirulence-Induced Gene and anthranilate synthase family are associated with bacterial response, genes from expanded families in poplar, covering α/β hydrolases, DUF567 proteins, and proteinase inhibitors, have been reported to be involved in response to fungal infection. Quantification of fungus-host distributions based on the fungal databases from the USDA Agricultural Research Service and literature (Lucas, 1998) reveals, for different regions worldwide, 1.5 to 106 times more fungal interactions for poplar compared with *Arabidopsis*. These findings indicate a strong correlation between the wide distribution of poplar–fungal interactions and the adaptive expansion of specific responsive gene families.

In *Chlamydomonas*, both tandem and block duplicates exhibit a strong GO enrichment for the term "chromatin assembly or disassembly." Inspection of the gene families responsible for this GO enrichment revealed that the four major types of histones (H2A, H2B, H3, and H4) are included. When analyzing other plant genomes, we observed that the histone family expansions were specific for *Chlamydomonas*. Detailed analysis of these genes reveals that there are 28 clusters that are composed of at least three different core histones (Figure 4C). During the S-phase of the cell cycle, large amounts of histones need to be produced to pack the newly synthesized DNA. In order to increase histone protein abundance, gene duplication, as also observed in mammalian genomes, provides a biological alternative compared with increased rates of transcription (Graves et al., 1985; Tripputi et al., 1986; Allen et al., 1991). Apart from sufficient histone proteins in rapidly dividing cells, exact quantities also are required for correct nucleosome formation. The assembly of histones occurs in a highly coordinated fashion: two H3/H4 heterodimers will first form a tetramer that binds the newly synthesized DNA and subsequently the addition of two H2A/H2B dimers completes the histone bead (Parthun et al., 1996;

PERSPECTIVE

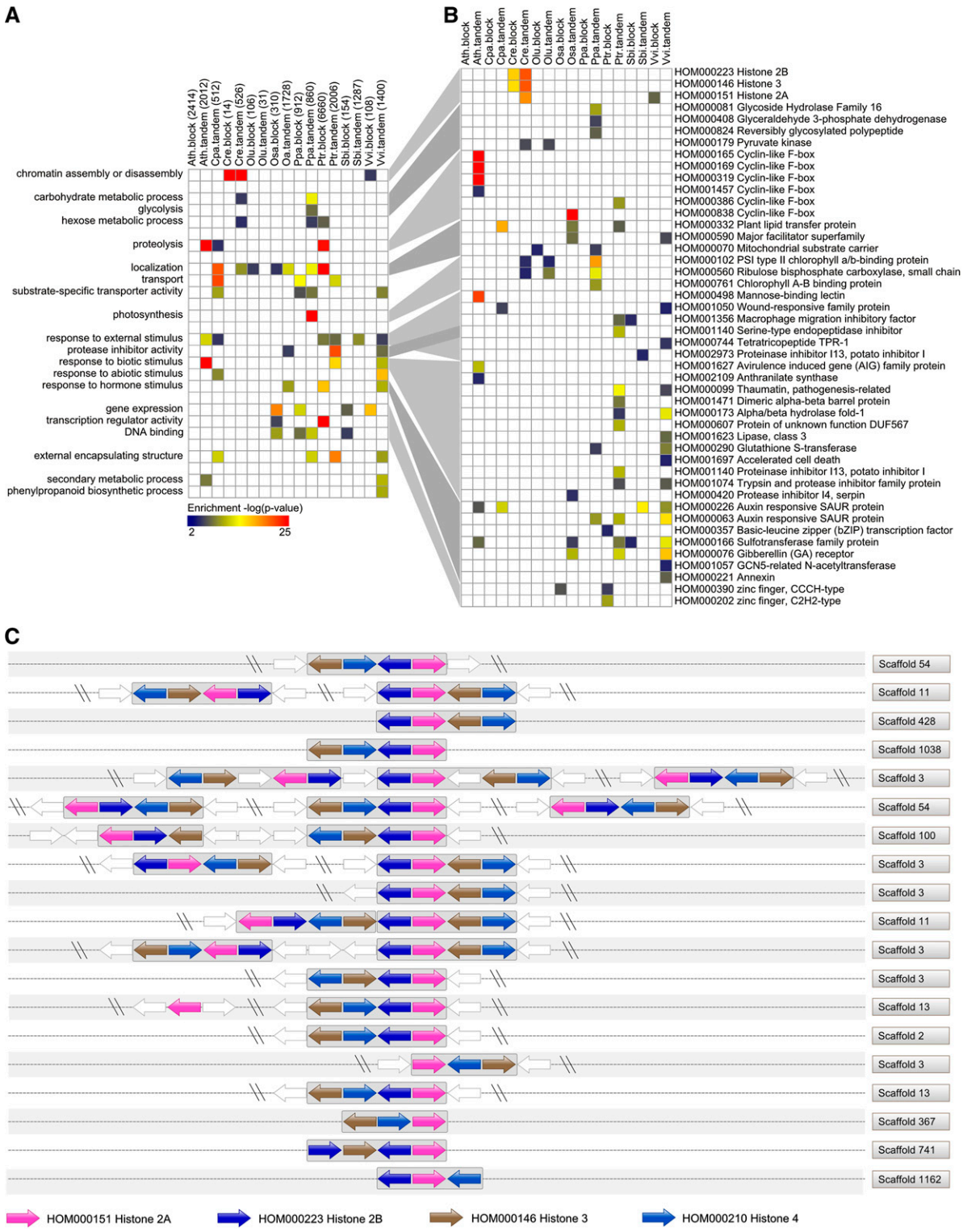

Figure 4. GO Enrichment Analysis of Species-Specific Gene Duplicates.

Grunstein, 1997). As shown in Figure 4C, the histone pairs that form dimers, which therefore should be present in equimolar amounts, occur very frequently in a divergent configuration (>95% of the histone genes occur in head-to-head pairs with their dimerization partner). This specific gene clustering suggests that bidirectional promoters guarantee equal transcription levels for the flanking genes (Fabry et al., 1995).

As a second case study, we used PLAZA to study large-scale duplication events in different lineages. Counting all gene duplication events for the different organisms confirms the presence of one or more WGD in *Arabidopsis*, moss, and monocots (see Supplemental Table 5 online). Interestingly, when analyzing the inferred ages of the different duplication nodes using the reconciled phylogenetic trees, we observed that the number of duplication events in the ancestor of angiosperms is larger than those in the eudicot ancestor (1880 and 1146 duplication nodes, respectively). In addition, these ancestral angiosperm duplications cover a larger number of gene families compared with the eudicot duplications (1141 and 757 families, respectively). This pattern suggests that, apart from the ancient hexaploidy detectable in all sequenced eudicot plant genomes (Tang et al., 2008b), older gene duplications have also significantly contributed to the expansion of the ancestral angiosperm proteome.

It is now generally accepted that, after the divergence of papaya and *Arabidopsis*, the latter species has undergone two rounds of WGD (Jaillon et al., 2007; Tang et al., 2008a; Van de Peer et al., 2009). PLAZA colinearity data were used to determine if levels of gene loss were different after the first (oldest) and second (youngest) WGD (also referred to as β and α, respectively). To this end, we selected multiplicons grouping four aligned *Arabidopsis* duplicated regions with an unduplicated outgroup region from either grape or papaya to count gene loss based on parsimony. Grapevine/papaya-*Arabidopsis* 1:4 alignments reveal that massive gene loss within *Arabidopsis* makes it very hard to link the homoeologous segments without aligning them to either grape or papaya (see Supplemental Figure 6 online) (Van de Peer et al., 2009). Manual inspection identified 26 reliable nonredundant multiplicons of which, in seven cases, the *Arabidopsis* segments could, based on K_s, unambiguously be grouped in two pairs that originated during the youngest duplication. All analyzed multiplicons can be visualized through the PLAZA website using a link reported in Supplemental Table 6 online. Analyzing all different patterns of gene loss using 139 ancestral loci (see Supplemental Table 6 online) revealed that 3.6 times more genes have been retained after the youngest α than after the oldest β *Arabidopsis*-specific WGD (31.13 and 8.63% retention, respectively). Consequently, this massive amount of gene loss masks most traces of the oldest WGD and explains why, with only the *Arabidopsis* genome available, the existence and timing of an older β duplication was debated (Simillion et al., 2002; Blanc et al., 2003; Bowers et al., 2003).

Comparison with Other Plant Genomics Platforms

The availability of online sequence databases and genome browsers provides an easy entry point for researchers to immediately investigate genome information without having to install any software. Furthermore, such services usually provide the possibility to link with an assembly of other Web-based resources (Brady and Provart, 2009). There has been a rapid growth in the number of plant genomics databases (Table 2). A major difference between these databases is the number of organisms included: whereas the Genome Cluster Database (Horan et al., 2005) and GreenPhylDB (Conte et al., 2008) only include *Arabidopsis* and rice, Gramene (Liang et al., 2008), PLAZA, and CoGe (Lyons and Freeling, 2008) have the most comprehensive set of species. CoGe includes, besides fully sequenced plant genomes, a large collection of viral, bacterial, fungal, and animal genomes. Comparing the data types, a noticeable trend is that most platforms focus on either gene families or genomic homology. Genome Cluster Database, GreenPhylDB, OrthologID (Chiu et al., 2006), and PlantTribes (Wall et al., 2008) all provide detailed information about gene families and phylogenetic trees but do not have any means to study genomic homology. By contrast, Plant Genome Duplication Database (Tang et al., 2008a), SynBrowse (Pan et al., 2005), and CoGe provide methods to study synteny and colinearity but do not include information about gene families. Phytozome (http://www.phytozome.net) and Gramene partially combine gene family and genome evolution data types. Whereas the former provides family-based local synteny plots, the colinearity framework in Gramene is based solely on genetic markers. Intraspecies dot plots are available in the Plant Genome Duplication Database, CoGe, and PLAZA and make it possible to investigate genes originating from WGD events. Finally, only Gramene, CoGe, and PLAZA provide a genome browser to obtain a general overview of a genomic region of interest.

Other platforms provide data focused on specific gene functions or sequence types but are not extensively described here.

Figure 4. (continued).

(A) The GO enrichment for species-specific block and tandem duplicates in different species is visualized using heat maps. Colors indicate the significance of the functional enrichment, while nonenriched cells are left blank. The number of genes per set is indicated in parentheses.

(B) Family enrichments indicate expanded gene families for different species. The gene sets are identical as in **(A)**. The gray bands link the enriched GO terms with the corresponding gene family expansions.

(C) The genomic organization of the core histone genes in *Chlamydomonas* reveals a pattern of dense clustering (indicated by gray boxes). Genes are shown as arrows; the direction indicates the transcriptional orientation and colors refer to the gene family a gene belongs to (families occurring only once are not colored for simplicity).

PERSPECTIVE

Table 2. Features of Plant Comparative Genomics Tools

Tool	Species[a]	Gene Families	Phylogenetic Trees	WGDotplots	Inter Species Colinearity	Functional Annotation	Genome Browser	Comments
PLAZA	9 (Ath, Cpa, Ptr, Vvi, Osa, Sbi, Ppa, Olu & Cre)	X	X	X	X	X	X	Multispecies colinearity views (Skyline Plot and Multiplicon view), K_s dating tool, family-wise similarity heat map and Workbench
Genome Cluster Database	2 (Ath & Osa)	X	X			X		Chromosome map and link with *Arabidopsis* expression data
GreenPhylDB	2 (Ath & Osa)	X	X			X		Manual curation of a subset of families
OrthologID	3+2 (Ath, Ptr & Osa + Ppa and Cre as outgroup)	X	X					Diagnostic characters per orthologous group
Plant Genome Duplication Database	7 (Ath, Cpa, Ptr, Mtr, Vvi, Osa & Sbi)			X	X			Genome-wide mapping tool for homologous sequences and syntenic locus search
Phytozome[b]	14 (Ath,Aly,Cpa, Ptr, Vvi, Mtr, Gma, Osa, Bdi, Sbi, Zma, Smo, Ppa & Cre)	X			±	X		
PlantTribes	5 (Ath, Cpa, Ptr, Mtr & Osa)	X				X		Link with *Arabidopsis* expression data
CoGe[c]	14 (Ath, Cpa, Ptr, Mtr, Lja, Vvi, Osa, Sbi, Zma, Ppa, Smo, Olu, Cre, Vca)			X	X	X	X	DNA-based sequence comparisons (conserved noncoding sequences)
SynBrowse	3 (Ath, Mtr, Lja)				X			Synteny browser based on GBrowse (no intraspecies colinearity)
Gramene[d]	6 (Ath, Osa, Ptr, Vvi, Sbi & Zma)	X	X	±	±	X	X	Based on the Ensembl pipeline

[a]Species names are abbreviated: *Arabidopsis lyrata* (Aly), *Arabidopsis thaliana* (Ath), *Brachypodium distachyon* (Bdi), *Carica papaya* (Cpa), *Chlamydomonas reinhardtii* (Cre), *Glycine max* (Gma), *Lotus japonica* (Lja), *Medicago trunculata* (Mtr), *Ostreococcus lucimarinus* (Olu), *Oryza sativa* (Osa), *Physcomitrella patens* (Ppa), *Populus trichocarpa* (ptr), *Sorghum bicolor* (Sbi), *Selaginella moellendorffi* (Smo), *Vitis vinifera* (Vvi), *Volvox carteri* (Vca), and *Zea mays* (Zma).

[b]Phytozome has a synteny viewer instead of a genuine colinearity pipeline.

[c]CoGe includes also viral, prokaryotic, and other, nonplant, eukaryotic genomes.

[d]Gramene has some features to visualize macrocolinearity based on marker maps.

Plant transcription factors can be studied using PlnTFDB (Riano-Pachon et al., 2007), AGRIS (Palaniswamy et al., 2006), and GRASSIUS (Yilmaz et al., 2009). The complementary platforms Phytome (Hartmann et al., 2006) and SPPG (Vandepoele and Van de Peer, 2005) are hybrid systems integrating gene information from genome sequencing projects with EST data for a comprehensive set of plant species.

Summary and Future Prospects

The PLAZA platform integrates genome information from a wide range of species within the green plant lineage and allows users to extract biological knowledge about gene functions and genome organization. Besides the availability of different comparative genomics data types, a set of interactive research tools, together with detailed documentation pages and tutorials, are accessible through a user-friendly website. Sequence similarity is used to assign protein-coding genes to homologous gene families, and phylogenetic trees allow the reliable identification of paralogs and orthologs. Through the integration of high confidence GO annotations and tree-based orthology between related plant species, we could (re-)annotate thousands of genes in multiple eudicot and monocot plants. Apart from local synteny plots that facilitate the identification of positional orthologs, gene-based colinearity is calculated between all chromosomes from all species and can be browsed using the so-called Skyline

plots. The WGDotplot visualizes all duplicated segments within one genome and dating based on synonymous substitutions generates an evolutionary framework to study large-scale duplication events. In addition, PLAZA's Workbench provides an easy access point to study user-defined gene sets or to process genes derived from high-throughput experiments. Based on a sequence similarity search or a list of gene identifiers, custom gene sets can rapidly be created and detailed information about functional annotations, associated gene families, genome-wide organization, or duplication events can be extracted. Consequently, this tool opens perspectives for researchers generating EST libraries from nonmodel species as these can easily be mapped onto a model organism. PLAZA hosts a diverse set of data types as well as an extensive set of tools to explore plant genome information (see Table 2).

Future efforts will be made to extend the number of available plant species and to include novel types of data to further explore gene function and regulation. Newly published plant genomes will be added on a regular basis to enlarge the evolutionary scope of PLAZA. The availability of genome information from more closely related organisms (Weigel and Mott, 2009) will make it possible to explore the similarities and differences between species at the DNA level and to identify, for example, conserved *cis*-regulatory elements on a genome-wide scale. In conclusion, PLAZA will be a useful toolkit to aid plant researchers in the exploration of genome information through a comprehensive Web-based research environment.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Phylogenetic Trees for Some Gene Families with Erroneously Identified Subfamilies.

Supplemental Figure 2. GO Projection Using Eudicot and Monocot Orthologous Groups.

Supplemental Figure 3. Synteny Plot Showing Conserved Gene Content between Different Genomic Regions Flanking Homologous Genes.

Supplemental Figure 4. K_s Dating Tool.

Supplemental Figure 5. Whole Genome Mapping Tool.

Supplemental Figure 6. The Gene Order Alignment of a *Vitis* Region and Four Corresponding α/β WGD *Arabidopsis* Regions.

Supplemental Table 1. Summary of Gene Family Content.

Supplemental Table 2. Overview of Subfamilies for 129 Large Gene Families.

Supplemental Table 3. Gene Counts before and after GO Projection per Organism.

Supplemental Table 4. GO Depth for Primary and Projected GO Annotations (Biological Process).

Supplemental Table 5. Overview of Duplication Events Inferred through Phylogenetic Trees (for Homologous Gene Families).

Supplemental Table 6. Counting Gene Loss in *Arabidopsis* Segment Generated by the α and β Whole-Genome Duplication.

Supplemental Methods 1. Data Retrieval and Delineation of Gene Families.

Supplemental Methods 2. Comparison of OrthoMCL Clusters with Phylogenetic Trees.

Supplemental Methods 3. Alignments and Phylogenetic Trees.

Supplemental Methods 4. Functional Annotation.

Supplemental Methods 5. Detection of Colinearity.

Supplemental Methods 6. K_s Dating.

Supplemental References.

ACKNOWLEDGMENTS

We thank Thomas Abeel, Eric Bonnet, Francis Dierick, and Stéphane Rombauts for technical assistance and Tine Blomme, Stefanie De Bodt, Jeffrey Fawcett, Elisabeth Wischnitzki, Eric Lyons, and the reviewers for helpful suggestions about the platform and tutorials. We thank Martine De Cock for help preparing the manuscript. S.P. thanks the Institute for the Promotion of Innovation by Science and Technology in Flanders for a predoctoral fellowship. K.V. is a Postdoctoral Fellow of the Research Foundation–Flanders. This work was supported by European Union EU-FP6 Food Safety and Quality Contract FOOD-CT-2006-016214. This project is funded by the Research Foundation–Flanders and the Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet).

Received September 22, 2009; revised December 4, 2009; accepted December 10, 2009; published December 29, 2009.

REFERENCES

Allen, B.S., Stein, J.L., Stein, G.S., and Ostrer, H. (1991). Single-copy flanking sequences in human histone gene clusters map to chromosomes 1 and 6. Genomics **10:** 486–488.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796–815.

Ashburner, M., et al. (2000). Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. **25:** 25–29.

Blanc, G., Hokamp, K., and Wolfe, K.H. (2003). A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res. **13:** 137–144.

Blomme, T., Vandepoele, K., De Bodt, S., Simillion, C., Maere, S., and Van de Peer, Y. (2006). The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. **7:** R43.

Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433–438.

Brady, S.M., and Provart, N.J. (2009). Web-queryable large-scale data sets for hypothesis generation in plant biology. Plant Cell **21:** 1034–1051.

PERSPECTIVE

Chen, F., Mackey, A.J., Vermunt, J.K., and Roos, D.S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS One **2:** e383.

Chen, Z.J. (2007). Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. Annu. Rev. Plant Biol. **58:** 377–406.

Chiu, J.C., Lee, E.K., Egan, M.G., Sarkar, I.N., Coruzzi, G.M., and DeSalle, R. (2006). OrthologID: Automation of genome-scale ortholog identification within a parsimony framework. Bioinformatics **22:** 699–707.

Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: How duplicated genes find new functions. Nat. Rev. Genet. **9:** 938–950.

Conte, M.G., Gaillard, S., Lanau, N., Rouard, M., and Perin, C. (2008). GreenPhylDB: A database for plant comparative genomics. Nucleic Acids Res. **36:** D991–D998.

Edgar, R.C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:** 1792–1797.

Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. **30:** 1575–1584.

Fabry, S., Muller, K., Lindauer, A., Park, P.B., Cornelius, T., and Schmitt, R. (1995). The organization structure and regulatory elements of Chlamydomonas histone genes reveal features linking plant and animal genes. Curr. Genet. **28:** 333–345.

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication: Tandem, whole-genome, segmental, or by transposition. Annu. Rev. Plant Biol. **60:** 433–453.

Fulton, T.M., Van der Hoeven, R., Eannetta, N.T., and Tanksley, S.D. (2002). Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell **14:** 1457–1467.

Gotz, S., Garcia-Gomez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talon, M., Dopazo, J., and Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. **36:** 3420–3435.

Graves, R.A., Wellman, S.E., Chiu, I.M., and Marzluff, W.F. (1985). Differential expression of two clusters of mouse histone genes. J. Mol. Biol. **183:** 179–194.

Grunstein, M. (1997). Histone acetylation in chromatin structure and transcription. Nature **389:** 349–352.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. **52:** 696–704.

Gutierrez, R.A., Green, P.J., Keegstra, K., and Ohlrogge, J.B. (2004). Phylogenetic profiling of the Arabidopsis thaliana proteome: What proteins distinguish plants from other organisms? Genome Biol. **5:** R53.

Hahn, M.W. (2007). Bias in phylogenetic tree reconciliation methods: Implications for vertebrate genome evolution. Genome Biol. **8:** R141.

Hanada, K., Zhang, X., Borevitz, J.O., Li, W.H., and Shiu, S.H. (2007). A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. Genome Res. **17:** 632–640.

Hanada, K., Zou, C., Lehti-Shiu, M.D., Shinozaki, K., and Shiu, S.H. (2008). Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol. **148:** 993–1003.

Hartmann, S., Lu, D., Phillips, J., and Vision, T.J. (2006). Phytome: A platform for plant comparative genomics. Nucleic Acids Res. **34:** D724–D730.

Horan, K., Lauricha, J., Bailey-Serres, J., Raikhel, N., and Girke, T. (2005). Genome cluster database. A sequence family analysis platform for Arabidopsis and rice. Plant Physiol. **138:** 47–54.

Hubbard, T., et al. (2005). Ensembl 2005. Nucleic Acids Res. **33:** D447–D453.

Hunter, S., et al. (2009). InterPro: The integrative protein signature database. Nucleic Acids Res. **37:** D211–D215.

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. Nature **436:** 793–800.

Jaillon, O., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449:** 463–467.

Koonin, E.V. (2005). Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. **39:** 309–338.

Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. **13:** 2178–2189.

Liang, C., et al. (2008). Gramene: A growing plant comparative genomics resource. Nucleic Acids Res. **36:** D947–D953.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell **133:** 523–536.

Lucas, J.A. (1998). Plant Pathology and Plant Pathogens. Plant Pathology and Plant Pathogens, 3rd ed. (Oxford:Blackwell Science).

Lyons, E., and Freeling, M. (2008). How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J. **53:** 661–673.

Merchant, S.S., et al. (2007). The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science **318:** 245–250.

Palaniswamy, S.K., James, S., Sun, H., Lamb, R.S., Davuluri, R.V., and Grotewold, E. (2006). AGRIS and AtRegNet. A platform to link cis-regulatory elements and transcription factors into regulatory networks. Plant Physiol. **140:** 818–829.

Pan, X., Stein, L., and Brendel, V. (2005). SynBrowse: A synteny browser for comparative sequence analysis. Bioinformatics **21:** 3461–3468.

Parker, M.S., Mock, T., and Armbrust, E.V. (2008). Genomic insights into marine microalgae. Annu. Rev. Genet. **42:** 619–645.

Parthun, M.R., Widom, J., and Gottschling, D.E. (1996). The major cytoplasmic histone acetyltransferase in yeast: Links to chromatin replication and histone metabolism. Cell **87:** 85–94.

Paterson, A.H. (2006). Leafing through the genomes of our major crop plants: Strategies for capturing unique information. Nat. Rev. Genet. **7:** 174–184.

Pennisi, E. (2007). Genome sequencing. The greening of plant genomics. Science **317:** 317.

Rensing, S.A., et al. (2008). The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science **319:** 64–69.

Riano-Pachon, D.M., Ruzicic, S., Dreyer, I., and Mueller-Roeber, B. (2007). PlnTFDB: An integrative plant transcription factor database. BMC Bioinformatics **8:** 42.

Rosin, F.M., and Kramer, E.M. (2009). Old dogs, new tricks: Regulatory evolution in conserved genetic modules leads to novel morphologies in plants. Dev. Biol. **332:** 25–35.

Rudd, S. (2003). Expressed sequence tags: Alternative or complement to whole genome sequences? Trends Plant Sci. **8:** 321–329.

Simillion, C., Janssens, K., Sterck, L., and Van de Peer, Y. (2008). i-ADHoRe 2.0: An improved tool to detect degenerated genomic homology using genomic profiles. Bioinformatics **24:** 127–128.

PERSPECTIVE

Simillion, C., Vandepoele, K., and Van de Peer, Y. (2004). Recent developments in computational approaches for uncovering genomic homology. Bioessays **26:** 1225–1235.

Simillion, C., Vandepoele, K., Van Montagu, M.C., Zabeau, M., and Van de Peer, Y. (2002). The hidden duplication past of *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **99:** 13627–13632.

Smith, J.M., and Smith, N.H. (1996). Synonymous nucleotide divergence: What is "saturation"? Genetics **142:** 1033–1036.

Stark, A., et al. (2007). Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature **450:** 219–232.

Stein, L. (2001). Genome annotation: From sequence to biology. Nat. Rev. Genet. **2:** 493–503.

Tanay, A., Regev, A., and Shamir, R. (2005). Conservation and evolvability in regulatory networks: The evolution of ribosomal regulation in yeast. Proc. Natl. Acad. Sci. USA **102:** 7203–7208.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008b). Synteny and collinearity in plant genomes. Science **320:** 486–488.

Tang, H., Wang, X., Bowers, J.E., Ming, R., Alam, M., and Paterson, A.H. (2008a). Unraveling ancient hexaploidy through multiply aligned angiosperm gene maps. Genome Res. **18:** 1944–1954.

The Reference Genome Group of the Gene Ontology Consortium (2009). The Gene Ontology's Reference Genome Project: A unified framework for functional annotation across species. PLOS Comput. Biol. **5:** e1000431.

Tripputi, P., Emanuel, B.S., Croce, C.M., Green, L.G., Stein, G.S., and Stein, J.L. (1986). Human histone genes map to multiple chromosomes. Proc. Natl. Acad. Sci. USA **83:** 3185–3188.

Tsesmetzis, N., et al. (2008). *Arabidopsis* reactome: A foundation knowledgebase for plant systems biology. Plant Cell **20:** 1426–1436.

Tuskan, G.A., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313:** 1596–1604.

Van de Peer, Y., Fawcett, J.A., Proost, S., Sterck, L., and Vandepoele, K. (2009). The flowering world: A tale of duplications. Trends Plant Sci. **14:** 680–688.

Vandepoele, K., Simillion, C., and Van de Peer, Y. (2002). Detecting the undetectable: Uncovering duplicated segments in Arabidopsis by comparison with rice. Trends Genet. **18:** 606–608.

Vandepoele, K., and Van de Peer, Y. (2005). Exploring the plant transcriptome through phylogenetic profiling. Plant Physiol. **137:** 31–42.

Vernot, B., Stolzer, M., Goldman, A., and Durand, D. (2008). Reconciliation with non-binary species trees. J. Comput. Biol. **15:** 981–1006.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. **19:** 327–335.

Wall, P.K., Leebens-Mack, J., Muller, K.F., Field, D., Altman, N.S., and dePamphilis, C.W. (2008). PlantTribes: A gene and gene family resource for comparative genomics in plants. Nucleic Acids Res. **36:** D970–D976.

Weigel, D., and Mott, R. (2009). The 1001 genomes project for *Arabidopsis thaliana*. Genome Biol. **10:** 107.

Yilmaz, A., Nishiyama, M.Y., Jr., Fuentes, B.G., Souza, G.M., Janies, D., Gray, J., and Grotewold, E. (2009). GRASSIUS: A platform for comparative regulatory genomics across the grasses. Plant Physiol. **149:** 171–180.