The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics

Laurel Cooper[1], Austin Meier[1], Marie-Angélique Laporte[2], Justin L. Elser[1], Chris Mungall[3], Brandon T. Sinn[4], Dario Cavaliere[4], Seth Carbon[3], Nathan A. Dunn[3], Barry Smith[5], Botong Qu[6], Justin Preece[1], Eugene Zhang[6], Sinisa Todorovic[6], Georgios Gkoutos[7], John H. Doonan[8], Dennis W. Stevenson[4], Elizabeth Arnaud[2] and Pankaj Jaiswal[1,*]

[1]Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331, USA, [2]Bioversity International, Parc Scientifique Agropolis II, 34397 Montpellier Cedex 5, France, [3]Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, [4]New York Botanical Garden, Bronx, NY 10458-5126, USA, [5]Department of Philosophy, University at Buffalo, Buffalo, NY 14260, USA, [6]School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, USA, [7]Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, B15 2TT, UK and [8]National Plant Phenomics Centre, Institute of Biological, Environmental, and Rural Sciences, Aberystwyth University, Aberystwyth SY23 3DA, UK

ABSTRACT

The Planteome project (http://www.planteome.org) provides a suite of reference and species-specific ontologies for plants and annotations to genes and phenotypes. Ontologies serve as common standards for semantic integration of a large and growing corpus of plant genomics, phenomics and genetics data. The reference ontologies include the Plant Ontology, Plant Trait Ontology and the Plant Experimental Conditions Ontology developed by the Planteome project, along with the Gene Ontology, Chemical Entities of Biological Interest, Phenotype and Attribute Ontology, and others. The project also provides access to species-specific Crop Ontologies developed by various plant breeding and research communities from around the world. We provide integrated data on plant traits, phenotypes, and gene function and expression from 95 plant taxa, annotated with reference ontology terms. The Planteome project is developing a plant gene annotation platform; Planteome Noctua, to facilitate community engagement. All the Planteome ontologies are publicly available and are maintained at the Planteome GitHub site (https://github.com/Planteome) for sharing, tracking revisions and new requests. The annotated data are freely accessible from the ontology browser (http://browser.planteome.org/amigo) and our data repository.

INTRODUCTION

Recent estimates show that the global population is projected to reach 9.6 billion people in the next few decades (http://www.wri.org/blog/2013/12/global-food-challenge-explained-18-graphics), and this presents enormous challenges for worldwide food production. Both basic and applied plant biology research frameworks are generating enormous quantities of next-generation plant science data from the high-throughput characterization of genomes, transcriptomes, proteomes and phenotypes along with large-scale genetic screens such as genome-wide association studies. Although this large volume of data is available, plant biologists, geneticists and breeders face the challenge of how to leverage this information efficiently and effectively. Finding novel gene targets and markers may help improve current plant germplasm and create new varieties, thus, ultimately contributing to the yield and quality of crops and feedstock for the growing global population, while also protecting the earth's environment.

Traditionally, many crop breeding communities have maintained their own standards for data formats and descriptors, for example, in the form of species-specific trait dictionaries. These data have important uses when it comes to formulating and testing hypotheses and for comparative analysis. However, researchers often face challenges in sharing this data due to incompatibilities in data formats

Table 1. Planteome reference ontologies and vocabularies

Ontology name	Knowledge domain	Source URL
Plant Ontology (PO)	plant structures and developmental stages	http://browser.planteome.org/amigohttps://github.com/Planteome/plant-ontology
Plant Trait Ontology (TO)	plant traits	http://browser.planteome.org/amigohttps://github.com/Planteome/plant-trait-ontology
Plant Experimental Conditions Ontology (PECO)	treatments and growth conditions used in plant science experiments	http://browser.planteome.org/amigohttps://github.com/Planteome/plant-experimental-conditions-ontology
Gene Ontology (GO)	molecular functions, biological processes, cellular components	http://www.geneontology.org/
Phenotypic Qualities Ontology (PATO)	qualities and attributes	https://github.com/pato-ontology/pato
Chemical Entities of Biological Interest (ChEBI)	molecular entities of biological interest focusing on 'small' chemical compounds	https://www.ebi.ac.uk/chebi/
Evidence and Conclusion Ontology (ECO)	evidence types for supporting conclusions in scientific research	http://www.evidenceontology.org/
Planteome NCBI Taxonomy*	taxonomic hierarchy	https://github.com/Planteome/planteome-ncbi-taxonomy

*The Planteome NCBI Taxonomy is a 'slice' or portion of the NCBI taxonomy file, with only those terms needed to create annotation to plants in the Planteome database. It is converted to an OWL file for loading into the Planteome Database.
Planteome reference ontologies cover a range of knowledge domains and are used to annotate plant genomics and phenomics data.

and descriptions of observations and experimental setups provided by different communities. Community-wide data sharing and re-use requires the use of common data standards and ontologies that provide a semantic framework for data collection, annotation and comparative analysis.

The Planteome project provides a centralized web portal (www.planteome.org) which features a suite of interconnected reference ontologies (listed in Table 1) utilized by the plant biology community for the annotation of plant gene expression data, traits, phenotypes, genomes and germplasm, across 95 plant taxa. The portal also hosts eight species-specific Crop Ontologies (CO) describing traits and phenotype scoring standards being adopted by international breeding projects on maize (*Zea mays*), sweet potato (*Ipomoea batatas*), soybean (*Glycine max*), pigeon pea (*Cajanus cajan*), rice (*Oryza sativa*), cassava (*Manihot esculenta*), lentil (*Lens culinaris*) and wheat (*Triticum aestivum*). The CO adoption is an integral part of the Integrated Breeding Platform and tools developed by the Consultative Group on International Agriculture Research (http://www.cgiar.org/). All ontologies and annotated data are available from the project website and browser, and through our web services and Application Programming Interface (API) for integration with software tools designed for data collection and curation, genome annotation and analysis. In this publication, we introduce the Planteome resource and provide details on how it can be accessed and utilized. We also introduce the Planteome Noctua gene annotation tool for engaging the research community in the functional annotation of plant genes.

THE PLANTEOME DATABASE

The Planteome database is accessible from our web site (http://planteome.org/) designed in Drupal version 6. It features an ontology browser and faceted search options to access ontologies and ontology-based annotations of various bioentities. The ontology browser (http://browser.planteome.org/amigo) is a customized adoption of the AmiGO browser (1), developed by the Gene Ontology Con-

sortium. All data and ontologies are stored in a SOLR (http://lucene.apache.org/solr) index system that allows for full-text searches through the ontology browser. The schema and index files for the design of the data store are available at the GitHub repository https://github.com/Planteome/amigo. In the current Planteome 2.0 Release, the Planteome database provides access to reference (Table 1) and species-specific ontologies and approximately 2 million (M) bioentities, or data objects, including proteins, genes, RNA transcripts, gene models, germplasm and QTLs (Quantitative Trait Loci; Table 2). Often more than one ontology term from the same or multiple reference ontology classes are used for bioentity annotation; the 2 M entities have approximately 21 M annotations to date. A mirror site of the Planteome database (not the website) is also accessible from the CyVerse cyberinfrastructure (http://cyverse.planteome.org).

Reference ontologies and vocabularies for plant biology

Ontologies provided by the Planteome can be used to annotate descriptions (for example, in experimental logs or trials), of the events, processes, conditions and observations for a broad range of entities. Such annotations range from the molecular function of a gene, its localization, and cell-, tissue- or organ-specific expression, to the broader role in response to growth environments and treatments at the whole plant or population level. In the current Planteome Release (2.0), the Planteome database includes a collection of 51 874 ontology terms (excluding obsoletes) from a suite of reference ontologies for plants (Table 1). They include the Plant Ontology (PO) (2–7), Plant Trait Ontology (TO) (8,9) and Plant Experimental Conditions Ontology (PECO) (8), which are developed in-house by the Planteome project. An additional set of reference ontologies include those developed by the collaborating groups, but relevant for use in annotation of plant biology data. These are the Gene Ontology (GO) (10), the Phenotypic Qualities Ontology (PATO) (11), Chemical Entities of Biological Interest (ChEBI) (12),

Table 2. Planteome database contents by bioentity type and number of annotations

Bioentity type	# Unique bioentities	# Ontology annotations
protein	1 674 962	14 984 245
germplasm	161 858	4 091 394
gene_model	35 988	1 501 291
mRNA	59 442	307 995
gene	42 995	222 205
QTL	13 873	49 296
gene_product	10 072	44 905
Noncoding RNAs (tRNA, miRNA, snoRNA, rRNA, snRNA)	1475	4786
Total number of unique bioentities and annotations	2 000 665	21 206 117

Each bioentity (gene, QTL, germplasm, etc.) may have more than one ontology annotation from the reference ontologies.

the Evidence and Conclusion Ontology (ECO) (13) and the NCBI taxonomy (14), described in Table 1.

All component reference ontologies are members of the OBO library (http://www.obofoundry.org/), and follow the guiding principles (15) suggested by the OBO Foundry, (http://www.obofoundry.org/principles/fp-000-summary.html) that foster a cooperative, interoperable community of vocabularies in ways which maximize opportunities for sharing and reuse. They include requirements that the ontologies be open, have a clear, defined scope, share a common format, use term-to-term relations which are unambiguously defined and are developed through a collaborative process for use by multiple resources.

Each ontology class has a unique primary term name and an alphanumeric identifier (ID) that forms part of a universal resource identifier, for example, the TO class: *leaf color* (TO:0000326) (http://purl.obolibrary.org/obo/TO_0000326; note that throughout the manuscript, ontology term names are written in italics). Terms in the ontologies should also have human-readable text definitions, including a citation to the source of the definition and to the names of the curators responsible for the annotation (16). Terms may also have synonyms, such as alternate names used in different plant research communities, or other languages. The synonym types include exact, related, narrow and broad (16). For example, the classes in the *plant anatomical entity* branch of the PO provide synonyms in Japanese and Spanish (6,7). Definitions often come with comments which provide additional information or examples of usage. Here, we provide brief summaries of the ontologies created and maintained by the Planteome Project.

Plant Ontology. The PO was developed in response to the need for a standardized terminology to describe plant anatomy and developmental stages for use in the annotation of plant genomics data (2–7). The PO consists of two branches; (i) the *plant anatomical entity* branch that describes *plant structures*, including *whole plant*, *plant anatomical spaces* such as the *axil*, and *plant substances* such as *cutin* and (ii) the *plant structure development stage* (PSDS) branch, that describes the stages of plant growth and development, such as the *flowering stage* or the *plant embryo development stage*. Terms in the PSDS are mapped to other plant development scales which are species or clade specific, such as Boyes *et al.* (17) for Arabidopsis and the Biologische Bundesanstalt, Bundessortenamt und CHemische industrie scale, which covers many crops such as cereals and grapes

(18). Future work on the PSDS may include integrating the Plant Phenology Ontology (19), which describes the timing of plant life-cycle events.

Ontology terms from the PO can be used to describe the spatial and temporal attributes of the sample source in an experiment. For example, the anatomical part from which samples were extracted, or on the growth stage at which the gene, QTL or phenotype was observed in a plant or population. For example, the rice gene *SD1* (*Semidwarf1*) is expressed in the *primary shoot system* during the *stem elongation stage*. The PO is designed to be species neutral, and thus, terms from the PO can be used to describe all green plants (the Viridiplantae). Terms in the PO are linked to annotated data from a wide variety of plants, ranging from traditional model species such as *Arabidopsis thaliana* to the crop plants such as maize, rice, and wheat that feed the world's growing population. In the current Planteome Release 2.0, there are approximately 1.4 M annotations to anatomy terms in PO and 1.1 M to development stage terms.

Plant Trait Ontology. A plant trait is a measurable characteristic of a plant or plant population, while a plant phenotype is an observed qualitative or quantitative value of a corresponding trait. For example, the TO term, *leaf color* (TO:0000326), is a commonly evaluated trait in plants. This term is combined with a quality term, such as *yellow* (PATO:0000324), to describe the phenotype *leaf color yellow*. Similarly, the trait *plant height* can be scored qualitatively as dwarf, tall, and semi-dwarf, as well as quantitatively by recording the absolute values, for example; 110, 96 cm, etc. to determine the phenotype. The TO facilitates interoperability between systems sharing trait data by providing pre-composed descriptions of a wide range of plant traits. The TO was originally created to describe rice QTL traits (20) and was expanded concurrently with the PO to encompass all green plants (Viridiplantae). Many of the TO classes follow the Entity–Quality (E–Q) pattern (11) where entity classes are drawn from the PO, GO and ChEBI and quality classes from PATO. The TO encompasses nine broad, upper-level categories of plant traits: *biochemical trait, biological process trait, plant growth and development trait, plant morphology trait, quality trait, stature or vigor trait, sterility or fertility trait, stress trait and yield trait*. Traits can be observed at any scale, ranging from molecular entities in plant cells, to cell types, tissues, organs, whole organisms and populations.

Plant experimental conditions ontology. PECO describes the biotic and abiotic treatments, growing conditions, and *study types* used in various types of plant biology experiments (8). For example, the conditions used to assess the response against a given type of treatment such as water deficit, red light, watering, photoperiod, soil type, fertilizer, nutrients, applications of growth hormones, and exposure to pest and pathogens.

Non-reference community vocabularies for plant biology

Crop ontology. In addition to the reference ontologies, the Planteome, includes a collection of species- or clade-specific application ontologies developed by the Crop Ontology (CO; http://www.cropontology.org/) project (21). The CO provides ontology-based descriptors for crop traits and standard variables for more than 20 crops to support field book design and phenotypic data annotations. To foster consistency in the data capture and annotation, each variable consists of a combination of a method and a scale suggested to be used for a given trait. Planteome curators work with CO developers and plant breeders to integrate terms used by their community by creating mappings to the reference ontologies, thus helping to connect phenotypes and germplasm annotations to genomics resources. Planteome Release 2.0 includes eight species-specific trait ontologies developed by the CO, for the crop plants, cassava (*Manihot esculenta*), maize (*Zea mays*), pigeon pea (*Cajanus cajan*), rice (*Oryza sativa*), sweet potato (*Ipomoea batatas*), soybean (*Glycine max*), wheat (*Triticum aestivum*) and lentil (*Lens culinaris*).

Planteome annotations

A key feature of the Planteome database is its use of the GO-derived strategy of annotations, described in detail by Hill *et al*. (22). An annotation is a link between an ontology term and a bioentity (also known as a data object). Annotations are created either manually by expert curators, or computationally and stored in Gene Association Format (GAF 2.0) files (http://www.geneontology.org/page/go-annotation-file-formats). The latter are essentially tab-delimited plain text files with the information organized into 17 columns. Each line in a GAF file corresponds to the assertion that some association exists between a bioentity and an ontology term. Each annotation includes a reference (usually to a PubMed ID) to support the assertion, and a conventional evidence code (http://planteome.org/evidence_codes). Statistics on the evidence types used to support the ontology annotations can be found in Table 3. This integration allows users to filter the annotation search results by the evidence type to an extent, by selecting the appropriate type from the facets available from the annotations results pages (e.g. http://browser.planteome.org/amigo/search/annotation). We are working on mapping these evidence codes to the reference Evidence and Conclusion Ontology (ECO), which will allow the users to find all or a subset of annotations supported by a given experimental evidence described by the ECO (Supplementary File SF1).

The files are stored in the Planteome Subversion Repository (http://planteome.org/svn/) and are loaded into the Planteome ontology browser, along with the network of ontologies. The curated data have been developed or sourced by Planteome curators and researchers at 20 other collaborating source databases (Table 4). The hyperlinked cross references in the browser connect the user to the original source for further information.

Terms from both branches of the PO and the other reference ontologies (Table 1) have been used to annotate mutant phenotypes in six plant species to facilitate cross-species querying for phenologs (orthologous phenotypes) and semantic similarity analyses (23). Classically, phenologs are defined as phenotypes related by the orthology of the associated genes in two species (24). In the Planteome, annotations may include any characteristic from molecular, functional, to gross level anatomical and growth stage observations. Therefore, Planteome annotations and the ontologies may help in answering questions such as 'Do the gene family members preserve similar phenotypes?' or 'Are the phenologs also true gene homologs?'

The annotation database is accessible online directly from the Annotation search page (see below) on the Planteome portal (http://browser.planteome.org/amigo/search/annotation), and the GAF files are also available for bulk download from the Planteome Subversion Repository (http://planteome.org/svn/).

Planteome ontology and data annotation browser

Researchers interested in exploring the Planteome database can access the ontologies and annotated data in various ways. The Planteome home page (http://planteome.org/) features a search box where you can search directly for ontology terms or bioentities. The menu has links to documentation, issue trackers, information on the project and its ontologies, publications and a contact form. To access the ontology browser (http://browser.planteome.org/amigo), users can click on the 'Ontology Browser' link on the Planteome home page and then click on the "Browse" button on the menu bar at the top. On the browser page (http://browser.planteome.org/amigo/dd_browse), users can explore the ontology hierarchy and associated annotation data using the 'drill down' browser (Figure 1A). From the top levels, one can open direct descendant terms individually by clicking on the + sign on the left hand side to create a custom view. Gray circles next to the ontology terms show the number of bioentities annotated to that ontology term. The filters on taxon and ontology/bioentity type on the left hand side (Figure 1B) allows further filtering of query results. By clicking on the gray circle, a popup window opens (Figure 1C) with term information including the identifier number, term name, definition, ontology source, synonyms, alternate IDs (if any) and synonyms. A hyperlink from the term name in the popup box will take the user to the ontology term detail page (Figure 1D). By selecting the 'Bioentities' link in the popup box, a new window opens with a list of all the bioentities associated with that term (Figure 1E).

Two search features are also available from the browser page. The 'Quick Search' box (Figure 1F) allows free text entries to be input and returns a list of related ontology terms and/or bioentities, and faceted searches can be performed using the 'Search' button (see below; Figure 1G).

Figure 1. An overview of the Planteome ontology and data annotation browser. (**A**) The drill-down browser allows users to explore the ontology hierarchy and the associated annotation data. Gray circles next to the ontology term names show the number of bioentities annotated to that ontology term either directly or accumulated indirectly from its children terms guided by the ontology tree and the term–term relationship types (**B**) Bioentities can be filtered by type and source taxon by selecting the red (exclude from search) or green (restrict search to) boxes on the left hand side. (**C**) Term information window appears if one clicks on an ontology term name and displays the alphanumeric identifier, term name, definition, ontology source, synonyms and alternate IDs (if any). (**D**) The term detail page can be accessed by clicking on the term name in the popup window, with additional information and links to direct and indirect annotations. (**E**) A full list of all the bioentities associated with the selected term can be opened by selecting the 'Retrieve Bioentities' link in the popup box. (**F**) Free text search box. (**G**) Faceted search menu.

Table 3. Total number of annotations present in the Planteome database supported by the respective evidence type

Evidence type	Annotations Total#
Inferred from Direct Assay (IDA)	3 146 811
Inferred from Expression Pattern (IEP)	2 351 012
Inferred from Experiment (EXP)	1 052 155
Inferred from Mutant Phenotype (IMP)	253 709
Inferred from Reviewed Computational Analysis (RCA)	37 244
Inferred by Curator (IC)	32 579
No biological Data available (ND)	26 240
Inferred by association of genotype from phenotype (IAGP)	19 030
Traceable Author Statement (TAS)	11 442
Inferred from Physical Interaction (IPI)	6706
Inferred from Genetic Interaction (IGI)	2715
Non-traceable Author Statement (NAS)	777
Inferred from Sequence Orthology (ISO)	13 436 322
Inferred from Electronic Annotation (IEA)	862 847
Inferred from Sequence Model (ISM)	28 013
Inferred from Sequence or structural Similarity (ISS)	25 536
Inferred from Biological aspect of Ancestor (IBA)	1079

The rows highlighted in light gray color are manual annotations and those in dark gray are results of automated/computational analysis. The project has integrated the use of ECO terms in its database by adopting the ECO-GO evidence code mappings (http://wiki.geneontology.org/index.php/Evidence_Code_Ontology_(ECO)) computationally, transforming the data to build associations to the ECO terms.

Table 4. Planteome database annotations by source database

Database source	URL	Annotations total#
Ensembl Plants	http://plants.ensembl.org/index.html	13 436 306
Genetic Resources Information Management System (GRIMS)	https://www.genesys-pgr.org/	2 978 495
Maize Genetics and Genomics Database (MaizeGDB)	https://www.maizegdb.org/	1 520 941
Germplasm Resources Information Network (GRIN)	https://www.ars-grin.gov/)	1 031 431
Planteome	http://planteome.org/	824 309
The Arabidopsis Information Resource (TAIR)	https://www.arabidopsis.org/	736 816
Gramene	http://www.gramene.org/	247 349
The *Physcomitrella patens* Resource (cosmoss)	http://www.cosmoss.org/	224 640
The Rice Annotation Project Database (RAP-DB)	http://rapdb.dna.affrc.go.jp/	73 952
The International Rice Informatics Consortium (IRIC)	http://iric.irri.org/home	54 727
Genome Database for Rosaceae (GDR)	https://www.rosaceae.org/	38 519
Sol Genomics Network (SGN)	https://solgenomics.net/	20 282
Jaiswal_lab	http://jaiswallab.cgrb.oregonstate.edu/	9561
Grape Genome Database (CRIBI_Vitis)	http://genomes.cribi.unipd.it/grape/	3410
SoyBase and the Soybean Breeders Toolbox	https://www.soybase.org/	2472
The European Arabidopsis Stock Centre (NASC)	http://arabidopsis.info/	1897
The Global Gateway to Genetic Resources (Genesys-pgr)	https://www.genesys-pgr.org/	389
AgBase	http://www.agbase.msstate.edu/	262
GenBank	https://www.ncbi.nlm.nih.gov/genbank/	210
Legume Information System (LIS)	https://legumeinfo.org/	146
National Center for Biotechnology Information (NCBI_gi)	https://www.ncbi.nlm.nih.gov/	3
Total number of annotations		21 206 117

Faceted searches. The 'Search' button (Figure 1G) on the menu bar opens the faceted search interface to query ontology terms, specific bioentities or annotated data. Searching for an ontology term results in a page listing the possible related terms (Figure 2A). Results can be filtered using the 'Ontology source' filter. If selected, it restricts the search results to the selected ontology. Other search options include filtering by subsets (terms that apply to a given taxa), ontology ancestors (moving up the tree) and whether or not the terms have been obsoleted. Users can go to an ontology term detail page by clicking on the hyperlinked term name in the results list.

The second faceted search option is a search for bioentities (data objects) in the Planteome database (Figure 2B). Options include filtering by source database, object type, taxon and direct and indirect (parent terms) annotated ontology terms. The third search option on the drop-down menu allows one to search for annotations between the
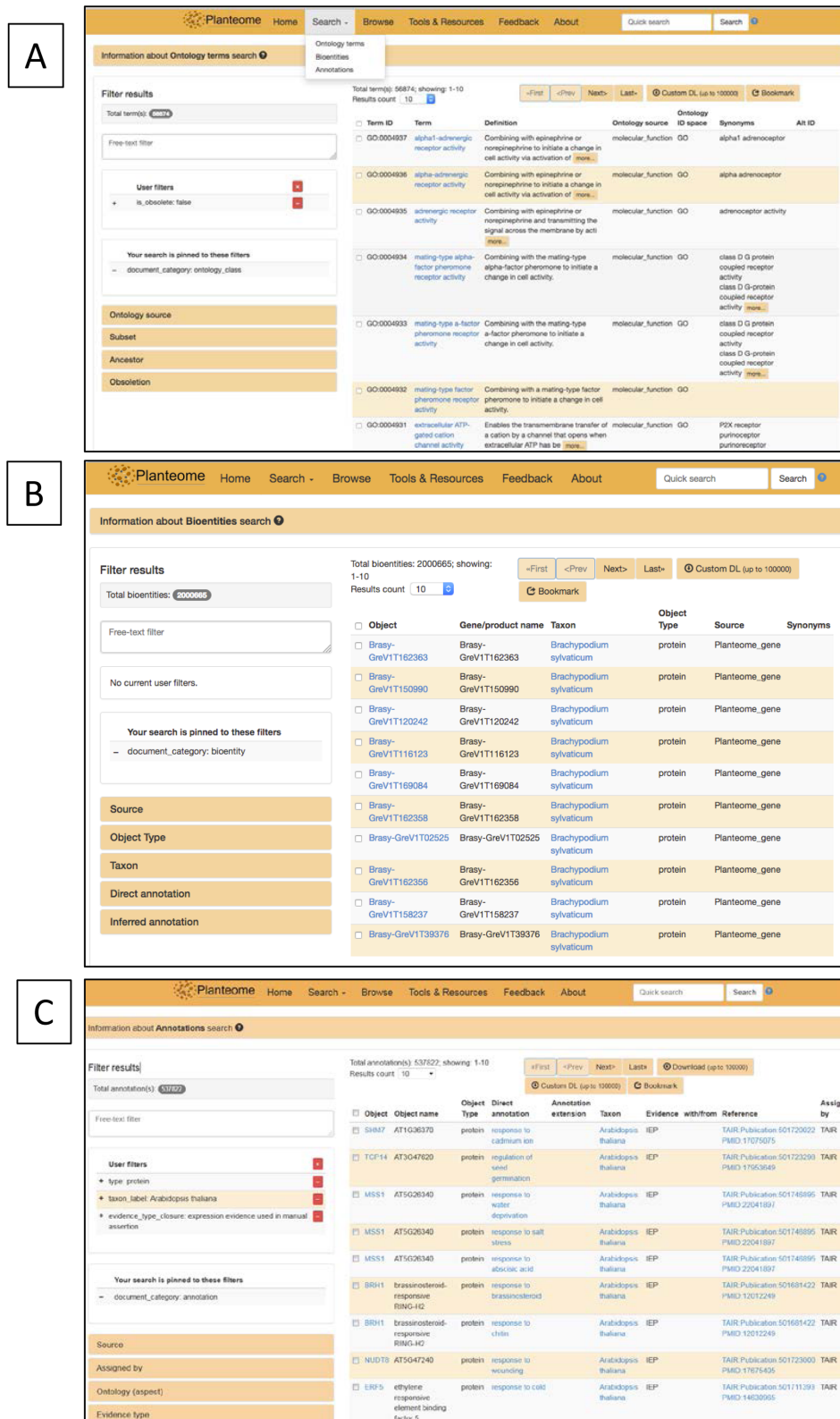
Figure 2. Faceted Searches for (**A**) Ontology terms, (**B**) Bioentities or (**C**) Annotations. Results can be filtered using the drop down menus on the left hand sides of each view.

bioentities and ontology terms (Figure 2C). On this page, the bioentities are listed, along with the associated ontology terms. Direct annotations are those made on the ontology term itself, while indirect annotations are those gathered on terms higher in the hierarchy of the ontology. The annotation search interface allows additional facets to filter the queries including evidence types, taxon, ontology (aspect), etc. (Figure 2C and Supplementary File SF1). On any of these three search pages, use the 'Free-text filter' box for further filtering. For example, performing a search for ontology terms, as in Figure 2A, enter the term 'leaf' and the results get filtered for only the terms that have the word 'leaf' in the name. Clicking the 'Bookmark' button on the window will generate a URL for safe keeping.

Data downloads

A custom download of up to 100 000 lines is allowed from any of the three faceted search pages shown in Figure 2. The interface allows selection of data fields for downloading in a tab-delimited, plain text file format. For advanced users, bulk annotations data files are freely available for download from the project SVN (http://planteome.org/svn) in the GAF format. The ontology files are accessible in OBO or OWL format from http://github.com/Planteome. The web services or API methods of data downloads and access are described later for advanced users.

Planteome community

Since the inception of the GO in 1998 (25), the global genomics community has come to recognize the importance of unified vocabularies to data interoperability. The plant genomics community has adopted the Planteome reference ontologies for use in a range of databases and genomics platforms (Table 5), from model organism sites such as TAIR and MaizeGDB to sites dealing with specific datasets such as protein domains (Superfamily), enzymes (BRENDA) or nuclear magnetic resonance spectra of metabolic profiles (MeRy-B). Several of the adopting projects (for example; Arapheno, BIP, Phenopsis DB, RARGEII and SGN) are using the Planteome ontologies to annotate and organize plant phenotyping data and a number of them, such as TAIR, MaizeGDB and SGN are contributing their annotated data to the Planteome Database.

Tools for collaboration and annotation

Web services. The Planteome project provides an API (http://planteome.org/web_services) that allows collaborators to access and use our data for intergation their web sites and applications. The API calls can be configured to query any of the ontology terms, their definitions, and other attributes, and annotation data, returning them in JSON format. The 'Search' method is fast enough to be used in an autocomplete search box, returning the basic information, while the 'Detailed Term Search' returns the complete data about a specific term. The Planteome project also delivers a standardized web service built on the BioLink API (http://biolink.planteome.org/api/). BioLink represents biomedical and biological entities

and the relationships between them. This includes genes, diseases, phenotypes, and metadata such as ontology terms, and is also used by projects such as the Monarch Initiative (https://monarchinitiative.org/) to drive portions of their website. The implemented BioLink API server (https://github.com/biolink/biolink-api) provides a swagger I/O end-point (https://swagger.io/) that can be used to automatically generate code to extract data in a uniform manner from a script or as part of an alternate web site. Using the exposed API we were able to provide a tool for the Galaxy Workflow Tool to expose Planteome data (https://toolshed.g2.bx.psu.edu/view/nathandunn/biolinkplanteome/66ece4fd024f), for example: http://biolink.planteome.org/api/bioentity/MaizeGDB%3A9024907/associations/?rows=20.

Ontology development and requests. All Planteome ontologies are maintained on the GitHub site (https://github.com/Planteome). The files for each Planteome reference and community ontology such as the COs are maintained in separate repositories. The ontology releases are managed through the GitHub release process. It allows collaborative development of ontologies by multiple registered and trained curators in various parts of the world. We use the issue tracker for each respective ontology repository for requesting new terms, edits or offering comments. For example, one can submit requests for PO terms at https://github.com/Planteome/plant-ontology/issues.

Gene annotation tool. Planteome Noctua (http://noctua.planteome.org/; Figure 3) is a web-based tool for collaborative curation and annotation of plant genes supported by empirical data and published literature sources. It is a customized version of the one used by the GO consortium (http://noctua.berkeleybop.org/). Registered users login with their GitHub credentials and can either create new annotations or edit existing ones. Planteome Noctua utilizes the reference ontologies described here including the GO; thus providing the ability to create a knowledge graph or model to annotate genes, or gene products and associate them with anatomical parts of a plant, developmental stages and/or traits, phenotypes, experimental conditions and treatments.

The core element of a statement in the model is called an 'individual' which can be an ontology term or a bioentity in the database. Individuals are linked together into units called 'annatons' by selecting the appropriate relationship and evidence type. For example, a curator may start by finding a bioentity (e.g. a gene) by filling in the gene name in the autocomplete box under 'Add Individual.' This can be repeated to create a list of ontology terms or bioentities. The appropriate relationships between the individuals are then created by connecting the blue dots and selecting the appropriate relationship type from the list. To add evidence to an individual, click on the empty circle in the box that illustrates a relation and an entity. In the resulting pop-up window, go to the 'Evidence' section and add an Evidence Type from the ECO, a supporting reference (either PMID, DOI or PO_REF), and where appropriate, an entry in the With/From field. These fields can be filled in the appropriate autocomplete boxes. Existing annotations can be

Table 5. Planteome ontologies are integrated into genomics platforms

Site name	link	comments
Annotare	https://www.ebi.ac.uk/fg/annotare/	Array Express experiment submission tool-samples may be tagged with PO terms during submission process
Arapheno	https://arapheno.1001genomes.org/ontology/	Database of Arabidopsis thaliana phenotypes can be browsed using TO or PECO
Arabidopsis Information Portal (Araport)	https://www.araport.org/	Search in ThaleMine by PO or GO terms
Brassica Information Portal (BIP)	https://bip.earlham.ac.uk/	TO and PO terms can be used to search population and trait scoring information related to the Brassica breeding community
BRENDA Enzyme Database	http://www.brenda-enzymes.info	PO, TO and EO are on their Ontology Explorer
Gramene Archive site	http://archive.gramene.org/plant_ontology/	Browse PO, TO, PECO and GO on the Ontology browser
Gramene Biomart	http://ensembl.gramene.org/biomart/martview/a5af63c60de7ebc805c5f558d7459deb	Can filter by PO and GO
Grape Genome Database Interface	http://genomes.cribi.unipd.it/cgi-bin/pqs2/query.pl?release=v1#Ontologies	Grape genes are annotated to PO terms
MAize Gene expressIon Compendium	http://bioinformatics.intec.ugent.be/magic/	Publicly available microarray data from Gene Expression Omnibus (GEO), and ArrayExpress, annotated with PO terms
MaizeGDB	http://maizegdb.org/gene_center/gene	PO annotations are listed on Gene pages
Maize Cell Genomics Database (MAGIC)	http://maize.jcvi.org/cellgenomics/index.php	Maize images tagged with PO terms
Manually Curated Database of Rice Proteins	http://www.genomeindia.org/biocuration/	Browse annotated data by PO, PECO, TO or GO term
Metabolomic Repository Bordeaux (MeRy-B)	http://services.cbib.u-bordeaux.fr/MERYB/vocabulary/ontology.php	Plant metabolomics platform database of Nuclear Magnetic Resonance metabolic profiles, browse by PO and PECO
The Compositae Genome Project (CGP)	http://compgenomics.ucdavis.edu/morphodb/analysis/viewOntology.php	Annotated data from lettuce and sunflower, can be browsed by PO hierarchy
Oryzabase PO site	http://shigen.nig.ac.jp/plantontology/ja/go	Japanese version of PO- plant structure terms translated to Japanese
Phenopsis DB	http://bioweb.supagro.inra.fr/phenopsis/	Arabidopsis Phenotype database, annotated with PO terms
Plant Ontology Enrichment Analysis Server (POEAS)	http://caps.ncbs.res.in/poeas/index.html	Plant phenomic analysis using PO terms based on genes from Arabidopsis thaliana
RIKEN Arabidopsis Genome Encyclopedia (RARGEII)	http://rarge-v2.psc.riken.jp/	Search mutant lines by phenotypes- use PO or PATO
Solanaceae Genome network (SGN)	https://solgenomics.net/tools/onto/index.pl	Browse and search by PO, GO, PATO terms
SuperFamily Browser	http://supfam.cs.bris.ac.uk/SUPERFAMILY/cgi-bin/phenotype.cgi?search=AP%3A0025099	Database of structural and functional annotation of protein domains and genomes- can browse by PO hierarchy
Virtual plant	http://virtualplant.bio.nyu.edu/cgi-bin/vpweb/	Browse the PO to see Arabidopsis genes annotated to that term
The Arabidopsis Information Resource (TAIR)	https://www.arabidopsis.org/	Browse by GO and PO ontology terms, annotation data can be downloaded
Wheat Data Interoperability Guidelines	http://ist.blogs.inra.fr/wdi/	Use of ontologies recommended by the Wheat Data Interoperability Working Group, of the Research Data Alliance
Cross Species Plant Phenotype Network	http://phenomebrowser.net/plant/	Results of the analysis conducted in the frame of the Plant Phenotype Pilot Project study with annotation files from six plant species (*Arabidopsis thaliana, Zea mays, Oryza sativa, Medicago truncatula, Glycine max* and *Solanum lycopersicum*)

imported from the Planteome ontology browser using the 'Function Companion' or 'GP Buddy' options by clicking on the green circle to edit annotations. Annotations created in Planteome Noctua can be downloaded in two different annotation file formats: GPAD and OWL, and can be converted to the GAF format for loading into the Planteome database. Our goal is to build Planteome as a common portal for collection, editing and distribution of the publicly annotated data on genes. Since Planteome Noctua allows use of multiple reference ontologies including GO, after integrating these annotations in the Planteome database, the gene annotations will be shared with our collaborators and the GO project by using both the APIs and the bulk downloads.

Data integration, curation and database development

The CO vocabularies are developed as species-specific, tab-delimited lists developed on the CO Trait Dictionary Version 5 format: (http://www.cropontology.org/CropOntology_Curation_Guidelines_20160510.pdf) and include traits, as well as the associated methods and scales of measurement. The CO trait terms (21) were mapped to equivalent or exact matching terms in the reference TO (26). Based on the mapping, the CO terms were added the TO graph as species-specific subclasses of their best matches
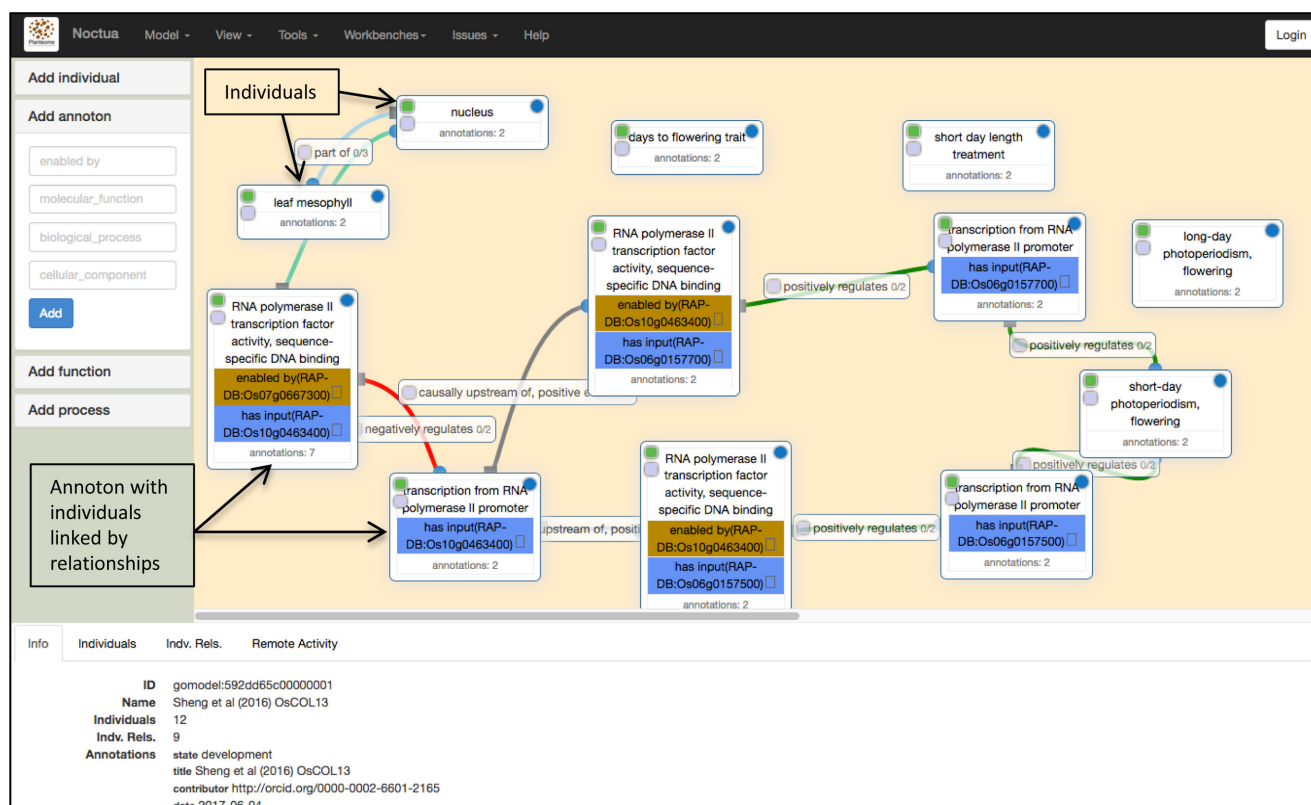
Figure 3. A view of a model under development in Planteome Noctua. Planteome Noctua (http://noctua.planteome.org/) is a web-based tool for collaborative curation and gene annotation supported by published literature or empirical data. Individuals from the reference ontologies are linked to one another through relationships and these assertions are supported by an evidence code from the Evidence and Conclusion Ontology. Once the model is complete, the information is saved and can be exported as a formatted file which can be processed to add the information to the database.

(Figure 4). It allows the CO to use the same species-neutral ontology tree from the TO as a starting point for building a robust ontology optimized for data sharing and integration between crop research communities. Thus avoiding the resources and time needed to create and duplicate the development of new species-specific ontologies. Additional species-specific ontologies can be easily created using only a flat list of species-specific traits, and their mappings to the reference TO. This allows for rapid development of application ontologies due to an existing foundation from which to build.

Standardized GO annotations for plant genomes and transcriptomes. Functional GO annotations were carried out in-house for 62 plant taxa. These annotations were done by integrating computational inferences from InterproScan (27) and projecting the manually curated annotations in Arabidopsis based on the orthology inferences driven by the InParanoid (28) clustering method described earlier (29,30). The orthology-based annotations were projected to the 62 species in a taxon-restrictive manner to avoid over projection and wrong annotations, e.g. flower development annotations from Arabidopsis were not projected to green algae. Duplicate annotations from the InterproScan and orthology-based annotations received higher confidence and were merged as unique. The Planteome is a unique annotation resource for finding annotations for many of the 62 species (http://planteome.org/node/128).

Germplasm annotations. A semi-automated pipeline was developed to create ontology-based annotations of plant germplasm (31). Many plant breeding and germplasm repository databases such as the USDA Germplasm Resources Information Network (GRIN: https://www.ars-grin.gov/) and The International Rice Informatics Consortium (IRIC: http://iric.irri.org/home) evaluate germplasm for a limited set of traits on their sites and record them in their databases using trait descriptors in plain text, a species-specific CO vocabulary or a proprietary controlled list. To improve interoperability of these data, a link between the individual trait descriptors and the reference ontology must be established. The native data format varies by source. To ensure proper data transformation and quality control before integration in the Planteome database, one of the first steps in annotating germplasm is mapping the source trait descriptors to the ontology terms from the reference TO. For example, 'pod color' trait evaluated in soybean/legume was mapped to the reference TO term *fruit color* (TO:0002617). This is followed by the data transformation step where a conversion script is run (script and examples available: https://github.com/Planteome/common-files-for-ref-ontologies/tree/master/scripts/germplasm_annotation) on
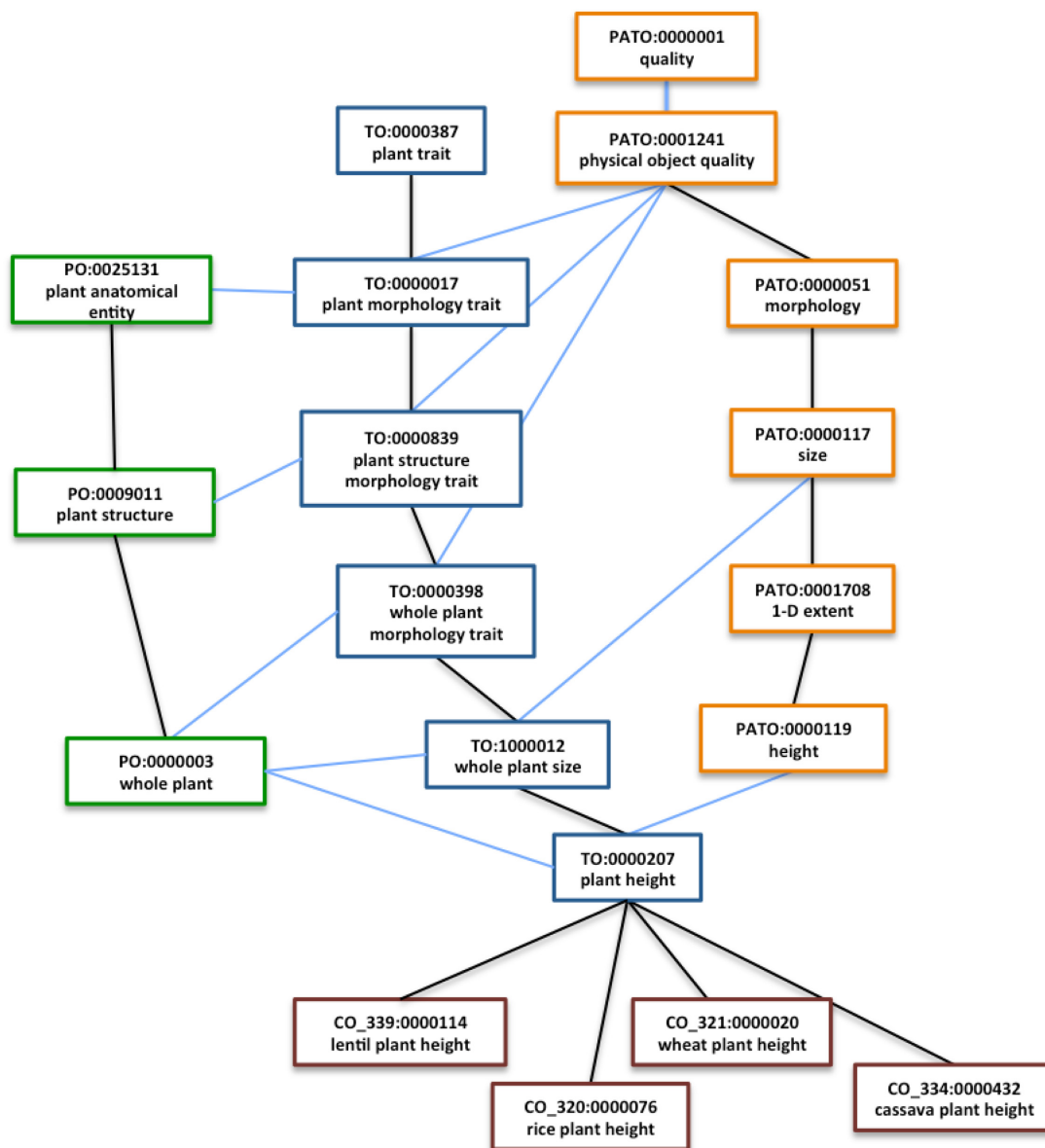
Figure 4. A view of the ontology hierarchy around Trait Ontology term plant height (TO:0000207). Crop Ontology (CO) terms for plant height from the lentil, wheat, rice and cassava ontologies are mapped to the Trait Ontology term for data integration.

the source data file and the trait mapping data to format the data files in the standard GAF 2.0 ontology annotation file format. The GAF formatted files are uploaded to the Planteome database at the time of the database build, and the resulting annotations provide hyperlinked cross references to the source. The original data must include three things: (i) a unique identifier for each germplasm entry, (ii) name of the evaluated trait and (iii) a phenotype score (observed qualitative/quantitative variables for the evaluated trait). It is important for the germplasm identifier to be unique in order to create a link back to the source database and avoid redundancies.

When available, we encourage providing additional useful pieces of information, such as, germplasm name synonyms, geographic location anme and GIS coordinates identifying the place where the original seed or plant was collected,

and where the phenotype was observed. In the current state, GAF formatted phenotype data is a must for integrating them in the Planteome database. The same GAF formatted files are also available to users for integration in their analyses and tools.

CONCLUSION AND FUTURE DIRECTIONS

The Planteome is a unique resource for both basic plant biology researchers such as evolutionary or molecular biologists and geneticists, and also for plant breeders who are interested in selecting for various traits of interest. The novel aspect of the Planteome lies in the semantic strength of the integrated ontology network, which can be traversed computationally. Planteome allows plant scientists in various fields to identify traits of interest, and locate data,

including germplasm, QTL and genes associated with a given trait, and can help in building hypotheses, confirming observations, data sharing and inter- and intra-specific comparisons. For example, plant biologists can use the annotation database to discover candidate genes from other species and compare the annotations to anatomy, growth stage and phenotypes supplemented with experimental evidence based on gene expression and analysis of mutants. Plant breeders on the other hand, are limited by the number of crosses they can make in a season and need to plan quickly. Therefore, tools built around the OMICs data and ontology-based annotations can help accelerate the process of identifying potential breeding targets, genetic markers, or previously evaluated germplasm with potential genetic underpinnings of agronomic traits, can help accelerate genetic gain by reducing downtime between genetic crosses. The ability to perform semantic queries on traits of interest is vital to this task. Also facilitated by the Planteome, one can identify germplasm and associated characters that would otherwise be housed in an obscure or poorly cross-referenced database and only tagged with free text descriptions, or unlinked vocabularies that represent a barrier to interoperability. The use of reference and species-specific ontologies for plants and the standardized annotations provided by the Planteome allows users the ability to leverage data from other studies and collaborate more efficiently.

Future directions for the Planteome project include the development of a reference Plant Stress Ontology and the addition of more species-specific vocabularies. We will be launching a plant gene nomenclature and annotation portal where researchers would be able to add new genes and annotations and edit existing ones. The data collected from these efforts will be shared semantically with sequencing projects, sequence archives and publishers of scientific literature for useful integration and consistency. Database user interface enhancements will include refinement of evidence and evidence code/ECO-driven faceted searches. The standardized functional annotation of the gene products will be further developed to assign plant Panther (32) gene family-based annotations and confidence scores in the projected annotations. We are also working on expanding the Planteome activities in the development of novel tools for, (i) automated recognition and ontology-based annotation of plant parts and phenotypes captured in plant images for taxonomic data collection (33,34), high throughput phenotyping projects, literature mining (35) and (ii) visualization of complex ontology trees and annotated data knowledge graphs in a user friendly manner.

AVAILABILITY

All Planteome project ontologies and source code are available in the Planteome project repositories on the GitHub https://github.com/Planteome. The annotation data in the standardized GAF2 file format are available for download from the Planteome Subversion (SVN) Repository (http://planteome.org/svn/).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. Carbon,S., Ireland,A., Mungall,C.J., Shu,S., Marshall,B., Lewis,S. and AmiGO Hub and Web Presence Working Group (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.
2. Jaiswal,P., Avraham,S., Ilic,K., Kellogg,E.A., McCouch,S., Pujar,A., Reiser,L., Rhee,S.Y., Sachs,M.M., Schaeffer,M. *et al.* (2005) Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. *Comp. Funct. Genomics*, **6**, 388–397.
3. Pujar,A., Jaiswal,P., Kellogg,E.A., Ilic,K., Vincent,L., Avraham,S., Stevens,P., Zapata,F., Reiser,L., Rhee,S.Y. *et al.* (2006) Whole-plant growth stage ontology for Angiosperms and its application in plant biology. *Plant Physiol.*, **142**, 414–428.
4. Ilic,K., Kellogg,E.A., Jaiswal,P., Zapata,F., Stevens,P.F., Vincent,L., Avraham,S., Reiser,L., Pujar,A., Sachs,M.M. *et al.* (2007) The Plant Structure Ontology, a unified vocabulary of anatomy and morphology of a flowering plant. *Plant Physiol.*, **143**, 587–599.
5. Avraham,S., Tung,C.-W., Ilic,K., Jaiswal,P., Kellogg,E.A., McCouch,S., Pujar,A., Reiser,L., Rhee,S.Y., Sachs,M.M. *et al.* (2008) The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Res.*, **36**, D449–D454.
6. Cooper,L., Walls,R.L., Elser,J., Gandolfo,M.A., Stevenson,D.W., Smith,B., Preece,J., Athreya,B., Mungall,C.J., Rensing,S. *et al.* (2013) The Plant Ontology as a tool for comparative plant anatomy and genomic analyses. *Plant Cell Physiol.*, **54**, e1.
7. Cooper,L. and Jaiswal,P. (2016) The plant ontology: a tool for plant genomics. In: Edwards,D (ed). *Plant Bioinformatics, Methods in Molecular Biology*. Springer, NY, Vol. **1374**, pp. 89–114.
8. Jaiswal,P., Ware,D., Ni,J., Chang,K., Zhao,W., Schmidt,S., Pan,X., Clark,K., Teytelman,L., Cartinhour,S. *et al.* (2002) Gramene: development and integration of trait and gene ontologies for rice. *Comp. Funct. Genomics*, **3**, 132–136.
9. Arnaud,E., Cooper,L., Shrestha,R., Menda,N., Nelson,R.T., Matteis,L., Skofic,M., Bastow,R., Jaiswal,P., Mueller,L. *et al.* (2012) Towards a reference Plant Trait Ontology for modeling knowledge of

plant traits and phenotypes. In: *Proceedings of the International Conference on Knowledge Engineering and Ontology Development*. SciTePress, Barcelona, Vol. **1**, pp. 220–225.

10. The Gene Ontology Consortium (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.*, **45**, D331–D338.

11. Gkoutos,G., Green,E., Mallon,A.-M., Hancock,J. and Davidson,D. (2004) Using ontologies to describe mouse phenotypes. *Genome Biol.*, **6**, R8.

12. Hastings,J., Owen,G., Dekker,A., Ennis,M., Kale,N., Muthukrishnan,V., Turner,S., Swainston,N., Mendes,P. and Steinbeck,C. (2016) ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.*, **44**, D1214–D1219.

13. Chibucos,M.C., Mungall,C.J., Balakrishnan,R., Christie,K.R., Huntley,R.P., White,O., Blake,J.A., Lewis,S.E. and Giglio,M. (2014) Standardized description of scientific evidence using the Evidence Ontology (ECO). *Database*, **2014**, bau075.

14. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.

15. Smith,B., Ashburner,M., Rosse,C., Bard,J., Bug,W., Ceusters,W., Goldberg,L.J., Eilbeck,K., Ireland,A., Mungall,C.J. *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–1255.

16. Walls,R.L., Athreya,B., Cooper,L., Elser,J., Gandolfo,M.A., Jaiswal,P., Mungall,C.J., Preece,J., Rensing,S., Smith,B. *et al.* (2012) Ontologies as integrative tools for plant science. *Am. J. Bot.*, **99**, 1263–1275.

17. Boyes,D.C., Zayed,A.M., Ascenzi,R., McCaskill,A.J., Hoffman,N.E., Davis,K.R. and Görlach,J. (2001) Growth stage–based phenotypic analysis of Arabidopsis A model for high throughput functional genomics in plants. *Plant Cell Online*, **13**, 1499–1510.

18. Hack,H., Bleiholder,H., Buhr,L., Meier,U., Schnock-Fricke,U., Weber,E. and Witzenberger,A. (1992) The extended BBCH-scale. *Allg. - Nachrichtenbl Deut Pflanzenschutzd*, **44**, 265–270.

19. Stucky,B., Deck,J., Denny,E., Guralnick,R.P., Walls,R.L. and Yost,J. (2016) The plant phenology ontology for phenological data integration. In: *Proceedings of the Joint International Conference on Biological Ontology and BioCreative*. CEUR Workshop Proceedings, Corvallis, CEUR-WS.org.

20. Ni,J., Pujar,A., Youens-Clark,K., Yap,I., Jaiswal,P., Tecle,I., Tung,C.-W., Ren,L., Spooner,W., Wei,X. *et al.* (2009) Gramene QTL database: development, content and applications. *Database*, **2009**, 1–13.

21. Shrestha,R., Matteis,L., Skofic,M., Portugal,A., McLaren,G., Hyman,G. and Arnaud,E. (2012) Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. *Front. Physiol.*, **3**, 326.

22. Hill,D.P., Smith,B., McAndrews-Hill,M.S. and Blake,J. (2008) Gene Ontology annotations: what they mean and where they come from. *BMC Bioinformatics*, **9**, S2.

23. Oellrich,A., Walls,R.L., Cannon,E.K., Cannon,S.B., Cooper,L., Gardiner,J., Gkoutos,G.V., Harper,L., He,M., Hoehndorf,R. *et al.* (2015) An ontology approach to comparative phenomics in plants. *Plant Methods*, **11**, 10.

24. McGary,K.L., Park,T.J., Woods,J.O., Cha,H.J., Wallingford,J.B. and Marcotte,E.M. (2010) Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 6544–6549.

25. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

26. Laporte,M.-A., Valette,L., Cooper,L., Mungall,C., Meier,A., Jaiswal,P. and Arnaud,E. (2016) Comparison of ontology mapping techniques to map plant trait ontologies. In: *Proceedings of the Joint International Conference on Biological Ontology and BioCreative*. CEUR Workshop Proceedings, Corvallis, CEUR-WS.org.

27. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.

28. Remm,M., Storm,C.E.V. and Sonnhammer,E.L.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons1. *J. Mol. Biol.*, **314**, 1041–1052.

29. Shulaev,V., Sargent,D.J., Crowhurst,R.N., Mockler,T.C., Folkerts,O., Delcher,A.L., Jaiswal,P., Mockaitis,K., Liston,A., Mane,S.P. *et al.* (2011) The genome of woodland strawberry (Fragaria vesca). *Nat. Genet.*, **43**, 109–116.

30. Myburg,A.A., Grattapaglia,D., Tuskan,G.A., Hellsten,U., Hayes,R.D., Grimwood,J., Jenkins,J., Lindquist,E., Tice,H., Bauer,D. *et al.* (2014) The genome of Eucalyptus grandis. *Nature*, **510**, 356–362.

31. Meier,A., Cooper,L., Laporte,M.A., Elser,J. and Jaiswal,P. (2016) Annotating germplasm to planteome reference ontologies. In: *Proceedings of the Joint International Conference on Biological Ontology and BioCreative*. CEUR Workshop Proceedings, Corvallis, CEUR-WS.org.

32. Mi,H., Huang,X., Muruganujan,A., Tang,H., Mills,C., Kang,D. and Thomas,P.D. (2017) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.*, **45**, D183–D189.

33. Lingutla,N.T., Preece,J., Todorovic,S., Cooper,L., Moore,L. and Jaiswal,P. (2014) AISO: annotation of image segments with ontologies. *J. Biomed. Semantics*, **5**, 50.

34. Kvilekval,K., Fedorov,D., Obara,B., Singh,A. and Manjunath,B.S. (2010) Bisque: a platform for bioimage analysis and management. *Bioinformatics*, **26**, 544–552.

35. Xu,W., Gupta,A., Jaiswal,P., Taylor,C. and Lockhart,P. (2016) Web application for extracting key domain information for scientific publications using ontology. In: *Proceedings of the International Conference on Biomedical Ontology and BioCreative (ICBO BioCreative 2016)*. CEUR Workshop Proceedings, Vol. **1747**, CEUR-WS.org.