The Planteome Project

Laurel Cooper, Austin Meier, Justin L. Elser, Justin Preece, Xu Xu, Ryan S. Kitchen, Botong Qu, Eugene Zhang, Sinisa Todorovic, Pankaj Jaiswal
Oregon State University, Corvallis, OR, USA

Marie-Angélique Laporte, Elizabeth Arnaud
Bioversity International, Montpellier, France

Seth Carbon, Chris Mungall
Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Barry Smith
University at Buffalo, Buffalo, NY, USA

Georgios Gkoutos
University of Birmingham, UK and University of Aberystwyth, UK

John Doonan
University of Aberystwyth, UK

Abstract— **The Planteome project is a centralized online plant informatics portal which provides semantic integration of widely diverse datasets with the goal of plant improvement. Traditional plant breeding methods for crop improvement may be combined with next-generation analysis methods and automated scoring of traits and phenotypes to develop improved varieties. The Planteome project (www.planteome.org) develops and hosts a suite of reference ontologies for plants associated with a growing corpus of genomics data. Data annotations linking phenotypes and germplasm to genomics resources are achieved by data transformation and mapping species-specific controlled vocabularies to the reference ontologies. Analysis and annotation tools are being developed to facilitate studies of plant traits, phenotypes, diseases, gene function and expression and genetic diversity data across a wide range of plant species. The project database and the online resources provide researchers tools to search and browse and access remotely via APIs for semantic integration in annotation tools and data repositories providing resources for plant biology, breeding, genomics and genetics.**

Keywords—ontology; traits phenotype; semantic; data integration, plants

I. INTRODUCTION

A. *Rationale*

It is estimated that the world population is projected to reach 9.6 billion people in next few decades (http://www.wri.org/blog/2013/12/global-food-challenge-explained-18-graphics). Therefore, the challenge is how to feed this growing population, while protecting the earth's environment. Traditional plant breeding methods for plant improvement may be combined with next-generation analysis methods, including the high-throughput and automated scoring of traits and phenotypes to develop improved varieties. Data from high-throughput sequencing, transcriptomic, proteomic, phenomic and genome annotation projects can be linked to germplasm resources through the use of interoperable, reference vocabularies (ontologies). In this way, the knowledge gained from the next-generation data can be utilized for crop improvement.

B. *What is the Planteome?*

The Planteome Project (www.planteome.org) is a centralized online informatics portal and database, consisting of a suite of reference ontologies for plants, an associated corpus of plant genomics and phenomics data, and tools for data analysis and annotation. Analyses of these data sets from genetic and genomic studies have the potential to improve our understanding of the molecular basis of economically relevant traits. In order to utilize this data, researchers must be able to connect the relevant plant traits of interest to the spatial and temporal expression patterns of genes, and elucidate their roles in biological processes in plants.

C. *Goals of the Planteome Project:*

1. A suite of interrelated reference ontologies to describe major knowledge domains of plant biology, comprising plant phenotype and traits, environments, and biotic and abiotic stresses.

2. Standards, workflows and tools for annotation of plant genomics data, and metadata for curation and improved annotation of genes, genomes, phenotype and germplasm.

3. The Planteome browser and database, a centralized, online informatics portal and repository where reference ontologies for plants are used to access data resources for plant traits, phenotypes, diseases, gene expression and genetic diversity data across a wide range of plant species.

4. Outreach involving the plant research community and K-12 and undergraduate students.

II. THE SCOPE OF THE PLANTEOME

The scope of the ontologies in the Planteome project ranges from a broad overview of plant environments and taxonomy, to the cellular and molecular level of expressed genes and their biological functions. The Planteome ontologies, described in more detail below, consist of the Plant Ontology (PO) [1-6], Plant Trait Ontology (TO) [7, 8], the Plant Environment Ontology (EO) [7] and the Plant Stress Ontology (PSO). The Planteome project imports and integrates with relevant reference ontologies developed by collaborating groups; the Gene Ontology (GO) [9, 10], the Phenotypic Qualities Ontology (PATO) [11], the Environment Ontology (ENVO) [12], and the Chemical Entities of Biological Interest (ChEBI) [13]. In addition, the Planteome integrates and maps species- or clade-specific application ontologies developed by the Crop Ontology (CO) project [14]. Together this suite of reference ontologies can be used to fully annotate and link together the vital plant knowledge domain.

The central reference ontology for plant anatomy and plant developmental stages, the Plant Ontology (PO) [1-6] grew out of the need to create associations between standardized terminology for plants and genomics data, and was based the work done to develop the Gene Ontology in the late 1990s [9, 10]. The PO is recognized worldwide as the reference ontology for plant structures and developmental stages, and is linked to data from a wide variety of plants, from traditional model species to the crop plants that feed the world's growing population.

Plant improvement relies on analyses of plant traits and phenotypes. For these purposes, the Plant Trait Ontology (TO) [9, 10] describes a wide range of precomposed plant traits consistent with Entity (E) - Quality (Q) statements and leads to an understanding of the molecular processes that underlie them. Each trait is a measurable or observable characteristic of a *plant structure* (PO:000901), a plant *cellular component* (GO:0005575), or a *plant structure development stage* (PO:0009012), as well as plant *biological processes* (GO:0008150) and *molecular functions* (GO:0003674). The TO encompasses nine broad, upper-level categories of plant traits: *biochemical trait* (TO:0000277), *biological process trait* (TO:0000283), *plant growth and development trait* (TO:0000357), *plant morphology trait* (TO:0000017), *quality trait* (TO:0000597), *stature or vigor trait* (TO:0000133), *sterility or fertility trait* (TO:0000392), *stress trait* (TO:0000164) and *yield trait* (TO:0000371).

The Plant Environment Ontology (EO) is used to describe the plant growth conditions and study types and can be combined with the terms from the other reference ontologies to fully annotate a plant phenotype description.

In addition to the reference ontologies, the Planteome works closely with developers of the species-specific vocabularies such as the Crop Ontology [14] to integrate their terms, create mappings to the reference ontologies and link phenotypes and germplasm to genomics resources.

III. DEVELOPMENT OF THE PLANTEOME ONTOLOGY NETWORK

The development of the Planteome Project ontology network is a fundamental change in the way of thinking about ontologies for plants. In the previous project, the Plant Ontology (http://www.plantontology.org/), a single reference ontology was developed and used to annotate plant genomic data to ontology terms describing plant structures and plant developmental stages. The addition of the other reference and species-specifc ontologies for plants enriches the annotation environment so a more complete picture of the metadata of plant pheotypes can be expressed.

In order to create the network, ontology terms in the TO and the species-specifc crop trait ontologies have been 'decomposed' into the corresponding Entity (E) - Quality (Q) statements which utilize terms from the other reference ontologies, such as PO and GO for the entities and PATO for the qualities. In this way, a network is formed which links all the various ontologies together.

One of the lessons learned in developing this network is that some of the reference ontologies and vocabularies developed by our collaborators (such as ChEBI, and the NCBI Taxonomy) are so large that they are cumbersome to display on our browser. For these, we have developed script to extract a relevant "slim" version which contains the needed terms.

IV. PLANTEOME ANNOTATION DATABASE

The Planteome database provides ontology terms and definitions along with the associated 'annotations' [15], between the ontology terms and data sourced from numerous plant genomics data sets. The Planteome 1.0 Beta Release (Nov. 2015) contains about 47 million annotations linking reference ontology terms to data objects representing genes, gene models, proteins, RNAs, germplasm and quantitative trait loci (QTLs) from 87 different plant species. These data are currently contributed by 29 different data sources. Planteome curators and researchers at various collaborating database groups work closely to develop the annotation files in the standardized data format database. The database is accessible online (http://planteome.org/) and also available for bulk download (http://palea.cgrb.oregonstate.edu/viewsvn/associations/).

The annotation database includes functional Gene Ontology annotations for 60 species. These predictions were done using two methods. The first method utilized an InterProScan [16] to identify protein domains. The resulting analysis files were then parsed to associate the protein domains to GO terms. The second method was to project ontology annotations based on

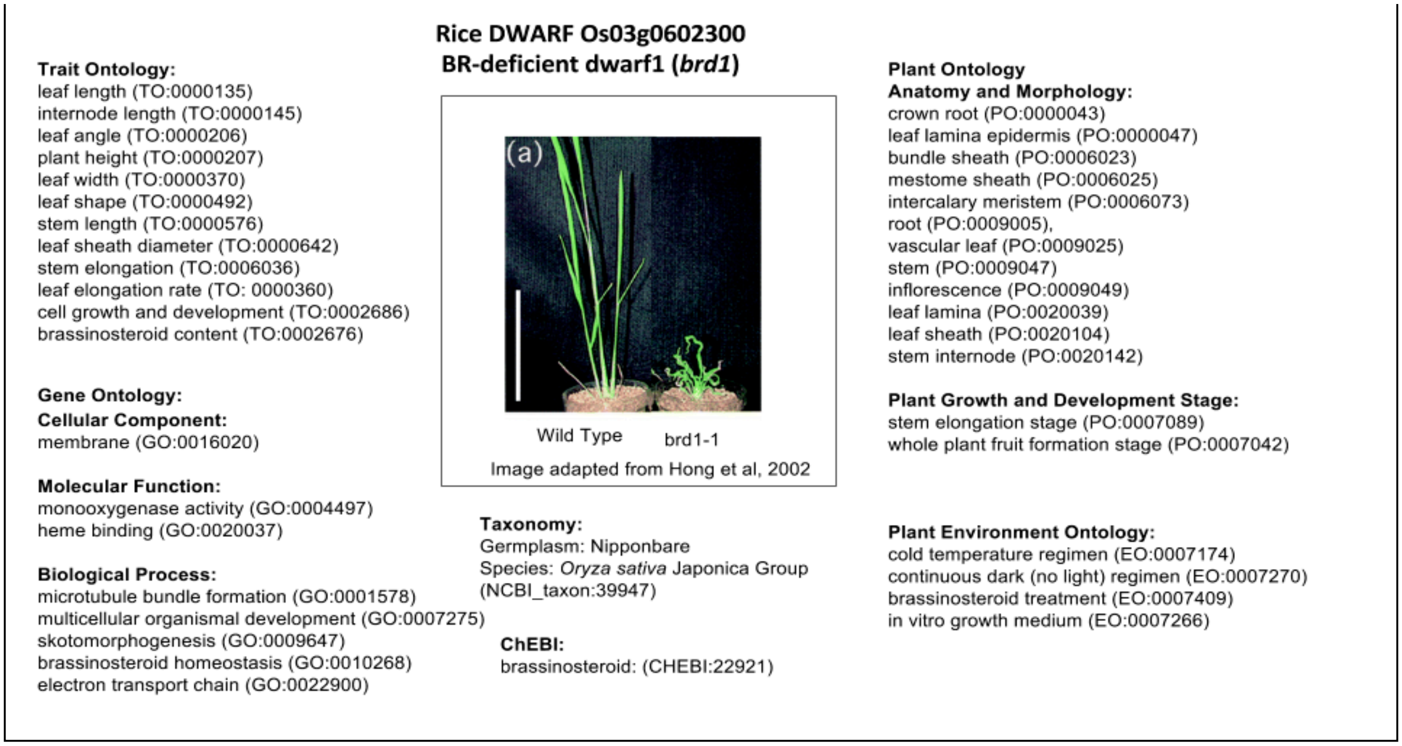
Rice DWARF Os03g0602300
BR-deficient dwarf1 (*brd1*)
Trait Ontology:
leaf length (TO:0000135)
internode length (TO:0000145)
leaf angle (TO:0000206)
plant height (TO:0000207)
leaf width (TO:0000370)
leaf shape (TO:0000492)
stem length (TO:0000576)
leaf sheath diameter (TO:0000642)
stem elongation (TO:0006036)
leaf elongation rate (TO: 0000360)
cell growth and development (TO:0002686)
brassinosteroid content (TO:0002676)
Gene Ontology:
Cellular Component:
membrane (GO:0016020)
Molecular Function:
monooxygenase activity (GO:0004497)
heme binding (GO:0020037)
Biological Process:
microtubule bundle formation (GO:0001578)
multicellular organismal development (GO:0007275)
skotomorphogenesis (GO:0009647)
brassinosteroid homeostasis (GO:0010268)
electron transport chain (GO:0022900)
(a)
Wild Type
brd1-1
Image adapted from Hong et al, 2002
Taxonomy:
Germplasm: Nipponbare
Species: *Oryza sativa* Japonica Group
(NCBI_taxon:39947)
ChEBI:
brassinosteroid: (CHEBI:22921)
Plant Ontology
Anatomy and Morphology:
crown root (PO:0000043)
leaf lamina epidermis (PO:0000047)
bundle sheath (PO:0006023)
mestome sheath (PO:0006025)
intercalary meristem (PO:0006073)
root (PO:0009005),
vascular leaf (PO:0009025)
stem (PO:0009047)
inflorescence (PO:0009049)
leaf lamina (PO:0020039)
leaf sheath (PO:0020104)
stem internode (PO:0020142)
Plant Growth and Development Stage:
stem elongation stage (PO:0007089)
whole plant fruit formation stage (PO:0007042)
Plant Environment Ontology:
cold temperature regimen (EO:0007174)
continuous dark (no light) regimen (EO:0007270)
brassinosteroid treatment (EO:0007409)
in vitro growth medium (EO:0007266)

Fig. 1. Annotation of Rice *brd1* mutant with reference ontology terms to capture the phenotype. The rice plant image is adapted with permission from [19] © John Wiley and Sons.

orthology to *Arabidopsis thaliana* genes. Orthology was predicted with InParanoid [17], a program that takes reciprocal BLAST output and uses pairwise similarity scores to determine orthologous clusters of genes. This is followed by creating gene super clusters by pooling species-pair clusters with common genes. The orthologous super clusters of the 60 species were compared with the known annotation files for *Arabidopsis thaliana* for GO, and new annotation files were generated. Planteome is the only online source providing GO functional annotation of genes identified for many of these species.

V. Case Study Example: Phenotype annotation of rice brassinosteroid (BR)-deficient dwarf mutant

Brassinosteroid (BR)-deficient (*brd1*) dwarf mutants of rice were characterized to determine the roles that BRs play in normal plant growth and development in a monocot plant [19]. Fig. 1 shows an example of how the reference ontologies can be used to annotate the phenotype of a (BR)-Deficient dwarf mutant rice, *brd1-1*. This image is a compliation of ontology terms from various Planteome reference ontologies that have been used to annotate the expression of *brd1* (Os03g0602300) in the Planteome database. These annotations were contributed from a variety of sources, such as Gramene (http://www.gramene.org/), EnsemblPlants (http://plants.ensembl.org/index.html), and The Rice Annotation Project (RAP) (http://rapdb.dna.affrc.go.jp/) and can be used to describe all aspects of the *brd1* mutant phenotype.

Gathering the annotations together in a unified platform such as the Planteome allows the data to be made accessible and facilitates gene discovery through inter- and intra-species comparisons.

VI. Planteome Tools for Collaboration and Ontology Integration

The Planteome project is developing a number of tools to increase access to the ontology terms and to increase the interoperability of the annotated data.

All the Planteome ontologies are publically available and are maintained at the Planteome GitHub site (https://github.com/Planteome) for sharing and tracking revsions. This site facilitates community feedback; users can make comments, request terms and suggest changes to the Planteome ontologies. In addition, the Planteome GitHub site also features species-specific vocabularies such as those from Crop Ontology (http://www.cropontology.org/).

Another new tool which is under development is a Trait Ontology-specific (http://to.termgenie.org/) instance of the TermGenie tool [20]. TermGenie uses a pattern-based approach to rapidly generate new terms and place them appropriately within the ontology structure. All terms are reviewed by a Planteome curator before the final commit to the ontology. TermGenie can be used to quickly obtain a TO term for annotation, if an appriopriate one does not already exist.

Planteome is developing an application programming interface (API) that will allow collaborators to access and use the hosted data in their web sites and applications. The first two API methods – currently accessible from the Planteome development environment – query Planteome-hosted ontologies for terms, term definitions, and other attributes, returning them in JSON format. The "search" method is fast enough to be used in an autocomplete search box.

All the Planteome reference and species-specific ontologies are available through the API service. Currently, the API only serves the term information, but the Planteome project plans to add API methods to access annotation data, as well.

The Planteome project is collaborating with the Bisque Image Analysis Environment (Center for Bio-Image Informatics, UCSB; http://www.cyverse.org/bisque) on integrated image segmentation and ontology annotation features. The Planteome project already hosts such a tool as a desktop application; Annotation of Image Segments with Ontologies (AISO; http://planteome.org/node/3), but we wish to move its functionality online as a module within Bisque, taking advantage of its shared CyVerse authentication, data store, and computation infrastructure. The ontology data itself will be served from external services, such as the Planteome API.

VII. Conclusions

The Planteome project is a centralized online plant informatics portal and which integrates reference ontologies for plants, and species-specific controlled vocabularies with a large and growing corpus of plant genomics data. This platform provides semantic integration of widely diverse datasets with the goal of plant improvement.

Acknowledgment

Funding for the Planteome project is provided by the National Science Foundation award IOS #1340112

References

[1] **Jaiswal, P, S Avraham, K Ilic, EA Kellogg, S McCouch, A Pujar, et al.,** 2005. Plant Ontology (PO): A Controlled Vocabulary of Plant Structures and Growth Stages. Comp Funct Genomics,. 6(7--8): p. 388-97 *(references)*

[2] **Pujar, A, P Jaiswal, EA Kellogg, K Ilic, L Vincent, S Avraham, et al.** 2006. Whole-plant growth stage ontology for angiosperms and its application in plant biology. Plant Physiol, 142(2): p. 414--28.

[3] **Ilic, K, EA Kellogg, P Jaiswal, F Zapata, PF Stevens, LP Vincent, et al.,** 2007. The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant. Plant Physiol. 143(2): p. 587--599.

[4] **Avraham, S, CW Tung, K Ilic, P Jaiswal, EA Kellogg, S McCouch, et al.,** 2008. The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. Nucleic Acids Res., 36(Database issue): p. D449--54..

[5] **Cooper L, Walls RL, Elser J, Gandolfo MA, Stevenson DW, Smith B, et al.** (2013) The Plant Ontology as a tool for comparative plant anatomy and genomic analyses. Plant and Cell Physiology **54**: e1–e1

[6] **Cooper L and Jaiswal P** (2016) The Plant Ontology: A Tool for Plant Genomics. *In* D Edwards, ed, Plant Bioinformatics. Springer New York, pp 89–114

[7] **Jaiswal P, Ware D, Ni J, Chang K, Zhao W, Schmidt S, et al.** (2002) Gramene: development and integration of trait and gene ontologies for rice. Comparative and Functional Genomics **3**: 132–136.

[8] **Arnaud E, Cooper L, Shrestha R, Menda N, Nelson RT, Matteis L, et al.** (2012) Towards a reference Plant Trait Ontology for modeling knowledge of plant traits and phenotypes. Proceedings of the International Conference on Knowledge Engineering and Ontology Development. Barcelona, Spain, pp 220–225.

[9] **Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. (2000)** Gene Ontology: tool for the unification of biology. Nat Genet 25: 25–29.

[10] **The Gene Ontology Consortium** (2014) Gene Ontology Consortium: going forward. Nucleic Acids Research. doi: 10.1093/nar/gku1179.

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

[12] **Buttigieg P, Morrison N, Smith B, Mungall C, Lewis S** (2013) The environment ontology: contextualising biological and biomedical entities. Journal of Biomedical Semantics 4: 43

[13] **Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, et al.** (2016) ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Research 44: D1214–D1219

[14] **Shrestha, R, Davenport, GF Bruskiewich, R, Arnaud**, **E.** (2011) Development of crop ontology for sharing crop phenotypic information. Drought phenotyping in crops: from theory to practice. pp 171–179

[15] **Hill DP, Smith B, McAndrews-Hill MS, Blake J (2008)** Gene Ontology annotations: what they mean and where they come from. BMC Bioinformatics 9: S2

[16] **Quevillon E, Silventoinen V, Pillai S, et al.** 2005. InterProScan: protein domains identifier. *Nucleic Acids Research*. 33(Web Server issue):W116-W120. doi:10.1093/nar/gki442.

[17] **Remm M, Storm CEV and Sonnhammer ELL** (2001). Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons. JMB, 314:1041-1052.

[18] **Altschul, SF, Madden, TL, Schäffer, AA, Zhang, J, Zhang, Z, Miller, W, et al.** (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

[19] **Hong, Z, Ueguchi-Tanaka, M, Shimizu-Sato, S, Inukai, Y, Fujioka, S, Shimada, Y, et al** (2002) Loss-of-function of a rice brassinosteroid biosynthetic enzyme, C-6 oxidase, prevents the organized arrangement and polar elongation of cells in the leaves and stem. The Plant Journal **32**: 495–508

[20] **Dietze, H, Berardini, T, Foulger, R, Hill, D, Lomax, J, OsumiSutherland, D, Roncaglia P, Mungall C** (2014) TermGenie - A web application for pattern-based ontology class generation. Journal of Biomedical Semantics **5**: 48

[21] **Lingutla N, Preece J, Todorovic S, Cooper L, Moore L, Jaiswal P** (2014) AISO: Annotation of Image Segments with Ontologies. Journal of Biomedical Semantics **5**: 50