Current Plant Biology 2 (2015) 1–11

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

The KnownLeaf literature curation system captures knowledge about *Arabidopsis* leaf growth and development and facilitates integrated data mining

Dóra Szakonyi[a,***], Sofie Van Landeghem[a,****], Katja Baerenfaller[b], Lieven Baeyens[a], Jonas Blomme[a], Rubén Casanova-Sáez[c], Stefanie De Bodt[a], David Esteve-Bruna[c], Fabio Fiorani[a,1], Nathalie Gonzalez[a], Jesper Grønlund[d], Richard G.H. Immink[e], Sara Jover-Gil[c], Asuka Kuwabara[b], Tamara Muñoz-Nortes[c], Aalt D.J. van Dijk[e], David Wilson-Sánchez[c], Vicky Buchanan-Wollaston[d], Gerco C. Angenent[e], Yves Van de Peer[a,f], Dirk Inzé[a], José Luis Micol[c], Wilhelm Gruissem[b], Sean Walsh[b,**], Pierre Hilson[a,g,*]

[a] *Department of Plant Systems Biology, VIB, and Department of Plant Biotechnology and Bioinformatics, Ghent University, B-9052 Ghent, Belgium*
[b] *Department of Biology, ETH Zurich, CH-8093 Zurich, Switzerland*
[c] *Instituto de Bioingeniería, Universidad Miguel Hernández, 03202 Elche, Alicante, Spain*
[d] *Warwick Systems Biology Centre, and School of Life Sciences, University of Warwick, Coventry CV4 7AL, United Kingdom*
[e] *Plant Research International, Bioscience, 6708 PB Wageningen, The Netherlands*
[f] *Genomics Research Institute (GRI), University of Pretoria, Private Bag X20, Pretoria 0028, South Africa*
[g] *INRA, UMR1318, and AgroParisTech, Institut Jean-Pierre Bourgin, RD10, F-78000 Versailles, France*

ARTICLE INFO

Article history:
Received 31 May 2014
Received in revised form 5 December 2014
Accepted 14 December 2014

Keywords:
Arabidopsis
Leaf growth
Literature curation
Data integration

ABSTRACT

The information that connects genotypes and phenotypes is essentially embedded in research articles written in natural language. To facilitate access to this knowledge, we constructed a framework for the curation of the scientific literature studying the molecular mechanisms that control leaf growth and development in *Arabidopsis thaliana* (*Arabidopsis*). Standard structured statements, called relations, were designed to capture diverse data types, including phenotypes and gene expression linked to genotype description, growth conditions, genetic and molecular interactions, and details about molecular entities. Relations were then annotated from the literature, defining the relevant terms according to standard biomedical ontologies. This curation process was supported by a dedicated graphical user interface, called Leaf Knowtator. A total of 283 primary research articles were curated by a community of annotators, yielding 9947 relations monitored for consistency and over 12,500 references to *Arabidopsis* genes. This information was converted into a relational database (KnownLeaf) and merged with other public *Arabidopsis* resources relative to transcriptional networks, protein–protein interaction, gene co-expression, and additional molecular annotations. Within KnownLeaf, leaf phenotype data can be searched together with molecular data originating either from this curation initiative or from external public resources. Finally, we built a network (LeafNet) with a portion of the KnownLeaf database content to graphically represent the leaf phenotype relations in a molecular context, offering an intuitive starting point for knowledge mining. Literature curation efforts such as ours provide high quality structured information accessible to computational analysis, and thereby to a wide range of applications.

[*] Corresponding author at: Institut Jean-Pierre Bourgin, INRA Centre de Versailles-Grignon, Route de St. Cyr (RD10), F-78026 Versailles Cedex, France. Tel.: +33 1 30 83 30 97; fax: +33 1 30 83 30 99.

[**] Corresponding author at: Albert-Ludwigs-University of Freiburg, Center for BioSystems Analysis, Faculty of Biology, Habsburgerstr. 49, D-79104 Freiburg, Germany

[***] Corresponding author at: Instituto Gulbenkian de Ciencia, Rua da Quinta Grande 6, 2780-156 Oeiras, Portugal. Tel.: +351 214464653.

[****] Corresponding author at: VIB-UGent, Department of Plant Systems Biology, Technologiepark 927, B-9052 Zwijnaarde, Belgium.

E-mail addresses: dszakonyi@igc.gulbenkian.pt (D. Szakonyi), sofie.van.landeghem@gmail.com (S. Van Landeghem), sean.walsh@biologie.uni-freiburg.de (S. Walsh), pierre.hilson@versailles.inra.fr (P. Hilson).

[1] Present address: Institute of Bio- and Geosciences, IBG2: Plant Sciences, Forschungszentrum Jülich, Wilhelm-Johnen-Straße, 52428 Jülich, Germany.

http://dx.doi.org/10.1016/j.cpb.2014.12.002

DATA: The presented work was performed in the framework of the AGRON-OMICS project (Arabidopsis GRO wth Network integrating OMICS technologies) supported by European Commission 6th Framework Programme project (Grant number LSHG-CT-2006-037704). This is a data integration and data sharing portal collecting all the all the major results from the consortium. All data presented in our paper is available here. https://agronomics.ethz.ch/.

1. Introduction

An increasing number of research papers are published every year containing vast amounts of scientific data. It is therefore difficult to rigorously monitor even a subset of these publications for a topic of particular interest. With the publication process becoming essentially digital, the scientific community now develops tools to capture aspects of published information into databases that can be queried by researchers to reveal previously veiled gene or protein functions [1].

Much effort has been devoted to the development of automatic text mining, a field supported by a growing community, with some initiatives focused on the processing of plant-related textual data [2–7]. However, most advanced text mining methods are not generic enough to be applied to a novel domain without additional work or to provide the level of detail required for specific studies.

A complimentary approach to text mining is the creation of high-quality datasets through the manual curation of primary research articles. In this context, specific nuggets of information are extracted with a rich and controlled vocabulary, compatible with algorithmic processing to identify valued relationships [8,9]. One of the most challenging tasks is to transform free text descriptions of complex phenotypes into structured statements linked to corresponding genotypes. Phenotype/genotype datasets are extremely valuable because they summarize current knowledge, they may reveal unknown biological mechanisms, and they facilitate the comparative analysis of functional studies across species [1,10,11].

As part of the AGRON-OMICS project, we created a standard framework to collect various types of information about *Arabidopsis* leaf growth (Fig. 1). Special attention was paid to gather phenotype data with the corresponding genotypes and additional parameters that characterize them as described in primary research papers. The components of our integrated system include (1) lists of selected ontology terms, (2) relationships between different types of entities expressed in a constrained structure, (3) a customized curation interface, (4) a semi-automated quality control pipeline, (5) a relational database capturing the collected data, and (6) an integrated network summarizing data curated within this project together with pre-existing knowledge. Information from 283 articles was compiled by multiple curators and the quality of the dataset was assessed for consistency. The workflow was designed in such a way as to fully support future text mining methods to be built on top of this data collection. All computer programs and data are publicly available at http://www.agronomics.ethz.ch in the "Knowtator, KnownLeaf, LeafNet" section.

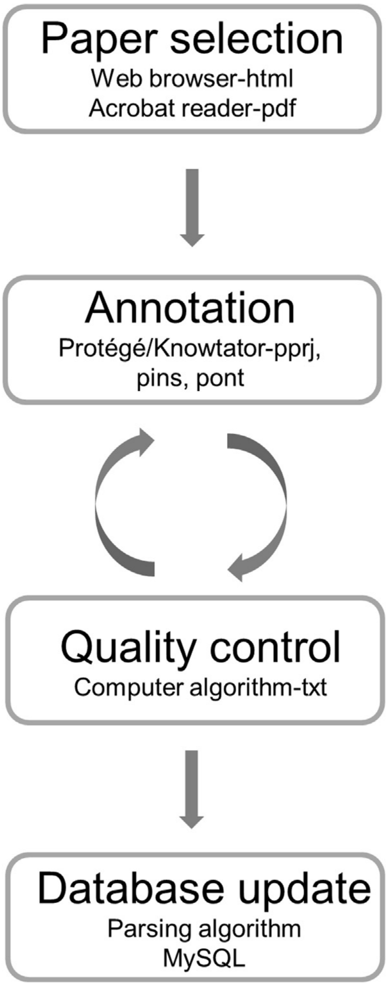


Fig. 1. Annotation workflow. Primary research papers were selected based on phenotypes, genes or interactions of interest. The full text html and pdf files were converted into a text format and imported into the Leaf Knowtator platform for manual annotation (.pont and .pins files contain domain classes and instances, respectively; the Protégé pprj project file identifies these files) The resulting data were exported from the Protégé software into individual XML files for each paper. These files were processed with dedicated in-house scripts for two distinct purposes. First, a quality control algorithm automatically detected predefined errors made during curation. Based on the output analytical text logs, the annotators corrected common inconsistencies within the Leaf Knowtator environment. The quality assessment and correction steps could be repeated until properly amended data files were obtained. Second, the corrected XMLs were parsed and flattened into the MySQL table embodying the KnownLeaf database.

2. Materials and methods

2.1. Annotation software and files

The Leaf Knowtator annotation interface was built with the software Protégé version 3.3.1 (http://protege.cim3.net/download/old-releases/3.3.1/full/) and the Knowtator plug-in version 1.9 beta (http://sourceforge.net/projects/knowtator/files/Knowtator/knowtator-1.9-beta2/). The program is available for various operating systems including Windows, Mac OS X, AIX, Solaris, Linux, HP-UX, any Unix platform and other Java-enabled platforms. Protégé requires Java 1.5 or a later version installed (http://www.java.com/en/download/index.jsp). Relations and corresponding slots were manually implemented with the Protégé/Knowtator tools. Ontology libraries were imported from publicly available onlice resources (Table S2).

The software was deployed as a standalone desktop application on each curator's computer and the Leaf Knowtator files (Leaf Knowtator.pprj, Leaf Knowtator.pins, Leaf Knowtator.pont) were shared. The annotations resulting from the curation of each individual paper were exported as an XML file from Protégé/Knowtator.

2.2. Databases

The MySQL database server (v5.x series) was used throughout together with the MySQL WorkBench graphical client tool. The annotations created by each curator were exported from Leaf Knowtator to produce one XML file per article, with a self-contained representation of the annotations consistent with the Protégé-Knowtator data-model, together with meta-data. Annotated phrases link to annotated classes, themselves compositions of slots associated with values (the ontology terms) through internally generated string identifiers. This complex network of classes and pointers was transformed with Perl ("knowtator2table.pl") into (1) tables in which each row represents a straight link between an annotated class, the corresponding text phrase and the assigned ontology terms or collection of terms, and (2) records to view each annotation as a collection of key value pairs. These tables and records were queried and viewed to identify missing or inconsistent annotations and to track progress.

The annotation tables were parsed into a MySQL relational database form with a Perl script ("PhenotypesEtc.pl"). The table **knowtator** represents the input tabular format with an auto-increment numeric row identifier ("id"). The table **knowtator_agi** holds references to AGI codes and has the id as a foreign key. The table **knowtator_papers** holds bibliographic information resolved by PubMed identifier with NCBI's efetch tool (http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi), together with the id as a foreign key, and the curator names per paper. In addition to data mining, this MySQL database served to generate summaries about content and to track the curation progress and the incremental growth of the data at a fine-grained level.

Additional publicly available sources of knowledge, such as gene and protein information (TAIR), protein–protein interactions, the *Arabidopsis* regulatory network (AGRIS AtRegNet) and the gene co-expression network (ATTED-II), were downloaded and parsed with Perl into a database format for uploading. These data and resources can be conveniently joined, filtered and summarized with structured query language (SQL). The database stored routines *user_get_knowtator_edges_for_cytoscape* and *user_get_knowtator_nodes_for_cytoscape* encode the filtering and joining functionality to develop the Cytoscape network files. Cytoscape version 2.8.3 was used throughout. The database and Perl scripts can be downloaded at www.agronomics.ethz.ch together with instructions on how to install the tables and data on an instance of the MySQL database server.

2.3. Quality control

The quality control script, applied to assist the expert curators to enhance data consistency and completeness, has been implemented as a Java (JDK 5) standalone program in Netbeans IDE 7.2. As input, the program receives a set of Knowtator-exported XML files, which are subsequently parsed and checked for completeness. Next, a quality check pipeline is run consisting of three different modules.

First, the program imposes the relation-slot structures as detailed in the annotation guidelines (Table S3), and reports any plausible violation of these guidelines, such as a missing required slot. These results are written to a human-readable log file, which can be opened with any text editor. Second, a quality module investigated the consistent usage of ontology terms. (i) The many text spans assigned to the same ontology term were listed together. In this way, the term "leaf_PO:0025034" was found to correctly relate to text spans such as "leaf" and "leaves", while the text fragment "rosette leaves" was reassigned to the more specific term "rosette leaf_PO:0000014". (ii) Instances where the same text referred to multiple ontology terms were also reported. For example, the word "irregular" was found to refer to the ontology term "abnormal" (PATO:0000460), but also "variant" (PATO:0001227). It may be acceptable to assign a different ontology term depending on context, but in most cases it is desirable to review synonymous annotations to increase consistency in the ontological assignments. (iii) Additional error logs reported the usage of undefined ontology terms as well as manually entered information not included in a relation (e.g. a highlighted gene name with no other linked data).

All quality logs produced by these fully automated scripts include the article identifier and the original textual information, enabling a fast look-up by the expert annotators in the Leaf Knowtator program. The annotator may choose to adjust the annotations or to ignore the log output when the annotations are deemed correct. After adjustments are made, the new XML file can be exported from Leaf Knowtator, and the process repeated until no more changes are necessary.

2.4. Training

The recruitment of community curators was supported with manuals and reference documents for the KnownLeaf annotation scheme and the Leaf Knowtator curation interface. It is recommended to first learn about the annotation scheme in the 'Annotation structures' document that provides a description of the ten relation categories and the corresponding table-like slot system. The 'Knowtator manual' is a hands-on guide to the Knowtator plug-in, including installation instructions, an introduction to the major curation functions, and general instructions on how to build an annotation project in the Knowtator framework. A 'Training document' with annotation solutions presents various examples that illustrate the different aspects of the curation process, useful for understanding the annotation practices before working on unannotated texts. All KnownLeaf project files can be downloaded at http://www.agronomics.ethz.ch, including the described annotation scheme, the original training document and a copy of the fully annotated training document.

3. Results

3.1. Data collection

Our first objective was to design and implement a framework to collect information about genes reported to be involved in leaf development (Fig. 1). We focused on phenotypes resulting from genetic alterations, but additional relevant relations were also captured, such as genetic interactions and protein–protein interactions (Table 1). Our annotation effort was focused on a specific subdomain of the available knowledge by imposing the following restrictions. (1) Data were acquired solely on the model organism *Arabidopsis thaliana*, because it offers the richest body of literature describing the molecular and genetic control of plant development. This initial choice does not preclude the later inclusion of articles describing gene functions and leaf phenotypes in other plant species, in particular major crops. (2) Statements were recorded if they referred to leaves, cotyledons, meristems, or the apical part of an embryo. (3) Text curation was limited to the Results sections of primary research articles, excluding the Introduction, Discussion and Supplemental data sections, to include actual data but avoid repetitions. For the same reasons, review articles were not taken into consideration.

The annotation of research articles was completed in two successive phases. The KnownLeaf system was initially developed on the basis of 174 publications curated by the reference annotator,

Table 1
Relation categories.

Relation	Example	References[a]
Phenotype	The rot3-2 allele causes enlarged leaf blades	Kim et al. (1999)
Gene expression	1-h BR treatments resulted in increased EXO transcript levels in. . .wild-type. . .plants	Coll-Garcia et al. (2004)
Feature	AtCPL2 contains one dsRNA-binding domain	Koiwa et al. (2002)
DNA–protein interaction	ARF2. . .bound to the promoter region of GH3.1	Wang et al. (2011)
Protein–protein interaction	AN3 interacted strongly with. . .AtGRF9	Horiguchi et al. (2005)
Genetic interaction	hyl1. . .appeared to suppress the as2 phenotypes	Xu et al. (2006)
Process	RHL2. . .involved during endocycles	Sugimoto-Shirasu et al. (2002)
Regulation of gene expression	AtCPL1. . .negative regulators of RD29A expression	Koiwa et al. (2002)
Regulation of process	AN3. . .promoting. . .cell proliferation	Horiguchi et al. (2005)
Regulation of phenotype	PHABULOSA. . .influence leaf shape	Garcia et al. (2006)

[a] Full references in Table S1.

to define the main principles of the process and to establish the annotation structure (Table S1). This initial set of articles was selected because they reported notable progress in the field of leaf growth and development, and described the function of key genes in relation with relevant mutant phenotypes. In the second phase, twelve community annotators, recruited within five laboratories part of the AGRON-OMICS consortium, were trained to work with the established tools, and curated an additional 109 articles, chosen because these were of particular interest to the respective contributors. The quality of the resulting composite data set was monitored through manual inspections, reiterative feedback between the reference and community annotators, and semi-automated tests as described below.

3.2. Leaf Knowtator, a custom-made curation tool

Powerful and adaptable tools are required to capture complex information and relationships involving leaf development in a structured and detailed framework. To this end, we chose the open-source software Protégé that was specifically created for ontology development and knowledge acquisition, and supported by an international user community [12]. Among the add-ons that expand the functions of this platform, the Knowtator plugin was designed to annotate text [13]. In combination, Protégé and Knowtator provide a flexible tool to build customized curation projects.

Ontology libraries or structured vocabularies were imported together with the hierarchical organization of their terms (Table S2). Ontologies and relations were entered as classes and each recorded statement defined a separate instance (according to Protégé definitions). Relations consisted of multiple slots in a tabular structure. Constraints (facets, in Protégé) were linked to the values contained in the slots.

Within the Leaf Knowtator interface (Fig. 2), the full text of the original research article is displayed in the center. There, the annotator selects a portion of the text that carries semantic knowledge and creates novel relation annotations accordingly by selecting from a list of relation categories. Each such relation is built up with a set of predefined information “slots” which appear automatically after selecting a relation type. These slots are filled in by the annotator who selects the relevant parts of the original text and tags them with an ontology term when appropriate. These textual fragments do not need to be contiguous nor originating from the same sentence, although this is most often the case. Additionally, the system automatically logs the name of the annotator, the date when the given relation was created, and the exact location of the tagged text span within the curated document.

As a result, both the original text and the attached ontology terms were linked together via the slot name. Often, the information implied in the original text was not clearly spelled out, in which case the appropriate ontology term was entered in the relevant slot without an explicit link to specific words in the original article. Alternatively, non-required slots could be left empty. A full example of a phenotype relation annotation is shown in Table 2 and its structure is explained in more detail in the next sections.

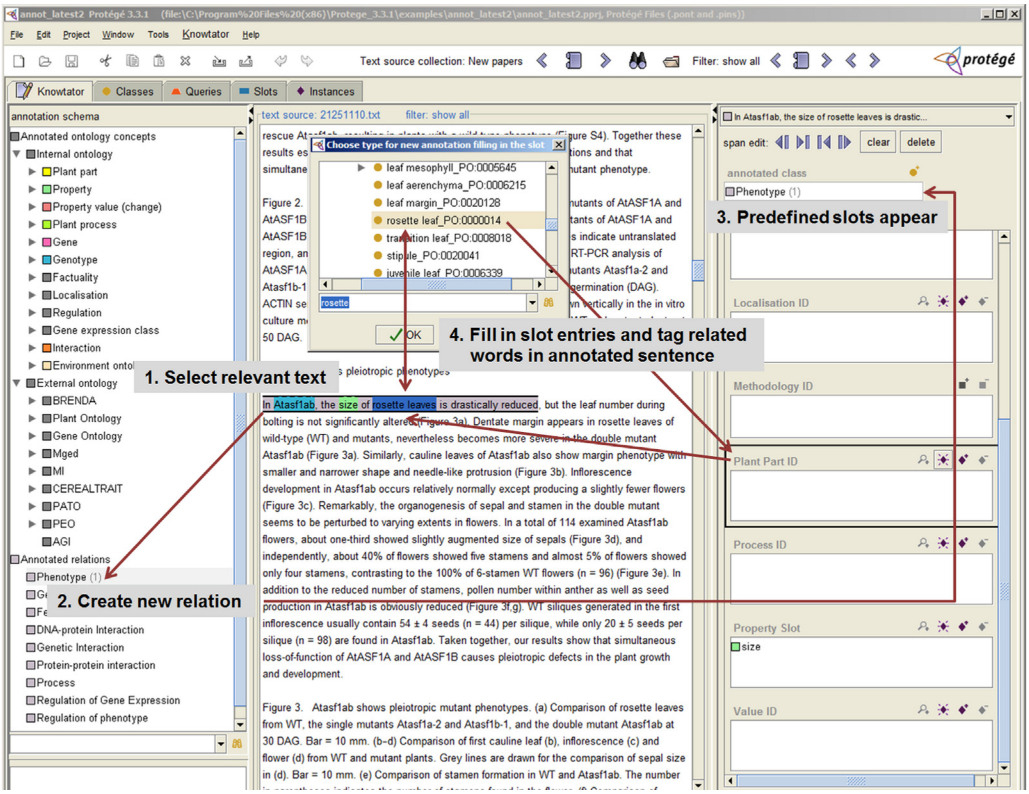
3.3. Definition of relation categories

Ten different categories of structured statements or “relations” were created to capture the data published on leaf growth and development (Tables 1 and 3). The relation names, in bold and italicized hereafter, are defined as follows. ***Phenotype*** entries contain morphological descriptions about wild-type, mutant or transgenic plants. ***Gene expression*** relations encode observations about RNA or protein levels corresponding to a given gene, in wild-type, mutant or transgenic background. They relate to molecular experiments such as (q)RT-PCR assays, microarray analyses, Northern and Western blots or capture information about gene expression patterns studied with a range of microscopy methods. The ***Feature*** category includes records of structural elements in DNA, RNA or protein molecules, e.g. promoter elements bound by transcription factors, miRNA target sites or protein domains. The ***DNA–protein***

Table 2
Example of a *Phenotype* annotation table. *Annotated sentence*: The reduced leaf area in the *hub1-1* mutant was confirmed by morphological measurements of the fully expanded leaves 1 and 2 [57].

Information type	Slot	Example	
		Annotated text	Entry
Genotype	Genotype	hub1-1	mutated gene_MI:0804
	Gene	hub1	RDO4 HUB1_AT2G44950
	Genotype Zygosity		homozygous diploid _APO:0000229
	Mutant LOF_GOF[a]		loss of function_APO:0000011
	Mutant type		
Phenotype	Plant part	leaf	leaf_PO:0025034
	Localization	leaves 1 and 2	juvenile leaf_PO:0006339
	Property	area	area_PATO:0001323
	Process		
	Value	reduced	decreased area_PATO:0002058
	Factuality		
	Developmental stage	fully expanded leaves	3 leaf fully expanded_PO:0001053
Environment	Growth condition		
Experiment	Methodology		

[a] LOF, loss of function; GOF, gain of function.



rescue At... ...ure S4). Together these results es... ...tions and that simultane... ...mutant phenotype.

Figure 2. ...mutants of AtASF1A and AtASF1B ...tants of AtASF1A and AtASF1B ...s indicate untranslated region, an... RT-PCR analysis of AtASF1A ...utants Atasf1a-2 and Atasf1b-1 ...germination (DAG). ACTIN se... ...vn vertically in the in vitro culture m... 50 DAG.

...s pleiotropic phenotypes

In Atasf1ab, the size of rosette leaves is drastically reduced, but the leaf number during bolting is not significantly altered (Figure 3a). Dentate margin appears in rosette leaves of wild-type (WT) and mutants, nevertheless becomes more severe in the double mutant Atasf1ab (Figure 3a). Similarly, cauline leaves of Atasf1ab also show margin phenotype with smaller and narrower shape and needle-like protrusion (Figure 3b). Inflorescence development in Atasf1ab occurs relatively normally except producing a slightly fewer flowers (Figure 3c). Remarkably, the organogenesis of sepal and stamen in the double mutant seems to be perturbed to varying extents in flowers. In a total of 114 examined Atasf1ab flowers, about one-third showed slightly augmented size of sepals (Figure 3d), and independently, about 40% of flowers showed five stamens and almost 5% of flowers showed only four stamens, contrasting to the 100% of 6-stamen WT flowers (n = 96) (Figure 3e). In addition to the reduced number of stamens, pollen number within anther as well as seed production in Atasf1ab is obviously reduced (Figure 3f,g). WT siliques generated in the first inflorescence usually contain 54 ± 4 seeds (n = 44) per silique, while only 20 ± 5 seeds per silique (n = 98) are found in Atasf1ab. Taken together, our results show that simultaneous loss-of-function of AtASF1A and AtASF1B causes pleiotropic defects in the plant growth and development.

Figure 3. Atasf1ab shows pleiotropic mutant phenotypes. (a) Comparison of rosette leaves from WT, the single mutants Atasf1a-2 and Atasf1b-1, and the double mutant Atasf1ab at 30 DAG. Bar = 10 mm. (b–d) Comparison of first cauline leaf (b), inflorescence (c) and flower (d) from WT and mutant plants. Grey lines are drawn for the comparison of sepal size in (d). Bar = 10 mm. (e) Comparison of stamen formation in WT and Atasf1ab. The number

Fig. 2. Leaf Knowtator interface. The Protégé/Knowtator interface consists of three main windows. The middle window displays the full text of a research paper. The relation categories and ontology collections are listed in the window on the left. Relation slots appear on the right side window as a series of smaller panels. When annotating an article, the curated sentence or sentence fragments are first selected and highlighted (gray background, black upper and lower lines) in the large middle window. To start recording a new statement, the relevant relation category is selected in the left window, resulting in the presentation of the corresponding slot panels. Each slot entry is then typed in or selected from ontology menus if applicable and tagged to the corresponding words in the middle window (color highlights).

Table 3
Summary of the KnownLeaf database content.

Relation category	# Annotations	# Unique AGI	Ratio
Phenotype	5608	381	14.7
Gene expression	4767	704	6.8
Genetic interaction	658	186	3.5
Feature	462	175	2.6
Protein–protein interaction	310	121	2.6
Process	235	140	1.7
Regulation of gene expression	204	70	2.9
Regulation of process	178	85	2.1
DNA–protein interaction	92	47	2.0
Regulation of phenotype	20	17	1.2
Total	12,534	883	14.2

interaction relations report the direct interaction between a protein and a target DNA molecule as established experimentally, for example with mobility shift or yeast one-hybrid assays, or chromatin immunoprecipitation. Similarly, ***Protein–protein interaction*** reports a direct interaction between two protein molecules, as determined by yeast two-hybrid assays, *in vitro* affinity enrichment experiments, co-immunoprecipitation, FRET assays, or split molecular tag studies. ***Genetic interactions*** report those relations. ***Process*** relations correspond to sentences with information about the biological or molecular function of a given gene or the corresponding gene product (RNA, protein). ***Regulation of gene expression*** reflects the functional activation or repression of genes rather than a direct mechanistic binding, which is comprised in the category ***DNA–protein interaction***. In this context, the regulation can take place at the DNA, RNA or protein level, for example via the action of transcription factors, epigenetic marks, small RNAs guiding transcript cleavage, or ubiquitin labels targeting proteins for degradation. The remaining two relation categories include general statements describing the involvement of genes or gene products in either biological processes, ***Regulation of process***, or phenotype, ***Regulation of phenotype*** without additional information. While not exhaustive, the factual results recorded with these ten distinct categories are sufficient for this scope.

3.4. Structure of the relations

Relations are defined in terms of specific parcels of data (“slots”), each containing a different type of information. A slot has a self-explanatory name such as ‘Plant part’ or ‘Growth condition’, can be linked to the corresponding words in the article text, and contains a single value to enable a seamless import of recorded relations into relational databases. In most cases, the data in a slot is a structured ontology term (see next section). However, some slots allow for free-text entries when a relevant ontology is not available or not detailed or extensive enough, such as the description of the experimental methodology. The structure of the ten relation categories is provided in Table S3. For clarity, slot names are hereafter italicized and underlined.

The principle of the annotation method is illustrated with a particular ***Phenotype*** relation in Table 2. The different slots (second column) of a relation can be logically grouped into a few high-level categories or “information units” such as Genotype, Phenotype and Environment (first column).

1. **Genotype**. This information unit consists of several slots specifying and identifying the relevant mutation and corresponding gene and mutant type. First, the required *Genotype* slot holds

one of the following values: (a) "Wild type_SO:0000817" when wild-type plants were studied (numbers here and below refer to the unique identifier in the listed ontology). (b) The "mutated gene_MI:0804" term for mutants with a known defective allele, irrespective of the exact nature of the mutation, except plants stably transformed with constructs that resulted in decreased gene expression that were labeled with the "knock down_MI:0789" term. (c) Overexpressor plants were labeled with the "over expressed level_MI:0506" term when they carried constructs for ectopic expression resulting in elevated RNA or protein levels. (d) A "Complex genotype" label captures transgenes or genetic configurations that cannot be described with any of the above terms, such as heterologous promoter-gene constructs or overexpressed transgenes in mutant backgrounds.

A *Gene* slot captures the appropriate AGI code for mutants, overexpressors and complex genotype entries, or a "no AGI" mark when there is no applicable code. Through the *Mutant LOF_GOF* slot, mutants can be further characterized by the nature of the mutation (e.g. LOF, "loss of function_APO:0000011" or GOF, "gain of function_APO:0000010") and the zygosity is stored in the *Genotype Zygosity* slot (e.g. "homozygous diploid_APO:0000229" or "heterozygous diploid_APO:0000230"). Finally, additional details about a mutation such as mutagen, allele, site of the lesion and exact change, can be recorded as free text in the slot *Mutant type*.

2. **Phenotype**. The phenotypic information unit is documented in a format reminiscent of the entity-attribute-value (EAV) model [14]. First, the entity is recorded in the *Plant part* slot (e.g. "leaf_PO:0025034"). Then, the attribute under consideration is subsequently filled into the *Property* slot (e.g. "area_PATO:0001323") to indicate the plant feature that was studied. Finally, the *Value* slot indicates the change of that specific feature (e.g. "decreased area_PATO:0002058").

   More detailed and flexible annotations can be introduced with a few additional slots. The *Localization* slot further specifies the plant part, for example to define in which organ a given cell type was located or when the subject is a subset of the original plant part term. The *Developmental stage* of the plant part can be recorded with the corresponding ontology term. GO terms can be entered in the *Process* slot. Finally, the *Factuality* slot qualifies certain statements with *ad hoc* labels such as 'negation' and 'speculation', to respectively mark negative statements or capture recorded statements that are suggested by the experimental data but not fully supported by additional evidence.
3. **Additional information types**. The details of plant growth conditions, special treatments or stress circumstances ("Environment") were entered in the *Growth condition* slot only if they were specifically stated in the Results section of the paper. These records relate to conditions that differ from the standard environment or have a direct effect on the phenotype. They included specific Plant Environment Ontology (EO) terms or in a few cases the CHEBI identifier of the chemical with which plants were treated. The *Methodology slot* (information type "Experiment") further allows for free text entries to describe the experimental method used.

The slots in the other nine relation categories are organized according to a similar scheme as detailed in the 'Annotation structure' section in Supplemental Information (Table S3).

3.5. Ontologies for standardized statements

Importantly, we adopted well-developed and widely accepted biological ontologies to build our annotations. The inclusion of such structured vocabularies enhances the interoperability of the resulting data, facilitating data integration and providing the proper basis for complex queries and computer reasoning. Concretely, whenever possible, the relevant words in the original articles were tagged with ontology terms from authoritative resources (Tables S2 and S4).

Plant organs, tissues and cell types were described with terms defined in the Plant Ontology (PO) and, in rare cases, in the BRENDA Tissue Ontology (BTO) [15,16]. Subcellular components were marked with Gene Ontology (GO) identifiers [17]. GO entries also provided information about biological processes. Plant traits and features, and changes that affect them, were described with terms from the Phenotype, Attribute and Trait Ontology (PATO) [12], Plant Trait Ontology (TO) [18], and BTO [15].

Relevant growth conditions or treatments were labeled with Plant Environment Ontology (EO) when specifically mentioned in the annotated Results section [18,19]. Unfortunately, the description of growth conditions and experimental treatments is generally not standardized [20,21], details describing experimental growth conditions vary widely between articles and are often cursory, and public ontologies addressing this semantic field are still under development. Therefore, it remains problematic to undertake coherent comparative or integrative analysis on the basis of fragmentary data, despite the basic relationship between phenotypes and the environment in which they are expressed [22,23]. Thus, exhaustive records of environmental conditions were not included.

Records about molecular events, modifications and interactions were defined with the Molecular Interaction (MI) standard assembled by the Proteomics Standards Initiative [19]. Finally, genes were mapped to their unique identifiers based on the *Arabidopsis* Genome Initiative (AGI) format and extracted from the latest TAIR10 genome annotation published by The *Arabidopsis* Information Resource [24]. While these common ontology libraries provided most necessary terms, they lacked some of the concepts important in leaf development biology. In such cases, we introduced *ad hoc* terms or included terms from alternative ontology systems, always ensuring consistency throughout the framework.

3.6. Monitoring relation consistency

Literature curation is prone to occasional errors and inconsistencies, especially when carried out by multiple contributors. To deliver a coherent data set, the annotation effort was designed with rigorous guidelines and community curators were trained during hands-on sessions backed up by documentation explaining the details of the annotation scheme and customized functions within Leaf Knowtator (see Supplemental Files, including an annotation manual and training documents). In addition, the quality of the records was monitored throughout the project with scripts designed to detect different types of errors that were subsequently corrected. Fully automated scripts first validated the completeness of the relation annotations, i.e. whether all required slots had been filled. Next, the consistency of ontology terms was automatically verified. Finally, orphan annotations or seemingly undefined ontology terms were also reported. Details of the quality control scripts are provided in Section 2.

The curators examined the resulting logs and the relations were adjusted when necessary. In rare cases, it was impossible to enter data in all required slots because textual descriptions were ambiguous or incomplete. While such annotations violated the initial guidelines, their information value often justified their inclusion in the final version of our database. As expected, the relations produced initially by the sole reference annotator – and main developer of Leaf Knowtator – were highly consistent and complete. Based on 174 curated articles, the quality control script reported on average only 3.1 missing required slots per article (19,267 required

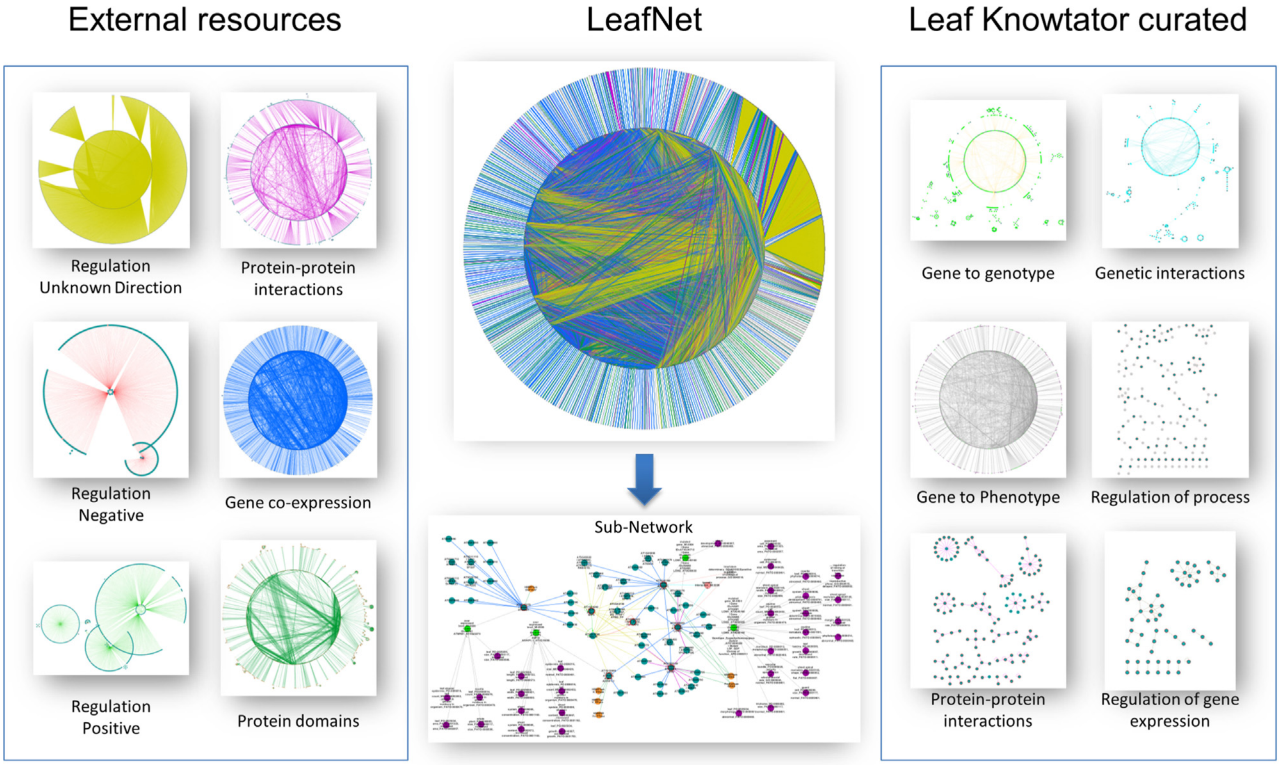


Fig. 3. Cytoscape representation of LeafNet. LeafNet is a composite assembly of publicly available knowledge resources (left panel) and Knowtator curated (right panel). The large central network represents the merged information from which sub-networks, including genes of interest, are derived using standard Cytoscape functionality. The sub-networks are practical to inspect the connectivity between genes of interest, the corresponding mutants and their phenotypes within the multi-faceted knowledge network landscape. See Fig. 4 for the details of the pictured subnetwork and the color code.

slots in total or an average of 111 slots per article; 2.8% missing slots). Note that this number is not expected to reach zero due to incomplete textual information. In contrast, the relations encoded originally by twelve community annotators contained on average 4.9 missing slots per article (369 missing slots among 12,122 slots, for 75 articles). After two rounds of automated evaluation with the quality control script followed by corrections, this average number dropped to 2.8, thus matching the quality of the initial set. Taking into account the total number of relations and corresponding mandatory slots, less than 2% of the required information could not be identified in the text.

3.7. The KnownLeaf database

XML files resulting from the curation of each annotated article were converted, flattened and parsed into database tables. The resulting KnownLeaf relational database consists of three tables: **knowtator** contains the annotated relations and has database foreign keys to **knowtator_papers** with bibliographic information and to **knowtator_agi** with AGI code references. The final database contains 9947 relations in the **knowtator** table with a total of 12,534 references to AGI codes, corresponding to 883 unique genes (3.2% of the TAIR10 protein coding genes) with on average 14.2 references per gene. Table 3 shows a breakdown of AGI references by annotated relation type in the completed data set. ***Phenotype*** statements form the largest category and refer to 381 distinct *Arabidopsis* genes, on average 14.7 statements per gene across the text corpus, highlighting the dense phenotype knowledge extracted from the literature. Statements on ***Gene expression*** are fewer but refer to more genes because the transcription of multiple genes was often described in a single mutant background or plant part.

A wide range of queries can be performed within the KnownLeaf database. For example, the annotated relations can be looked up for a given *Arabidopsis* gene or protein; genes can be searched that are linked to a particular plant part or to specific phenotype alterations (see example SQL within https://www.agronomics.ethz.ch/knowtator/code_repository.zip). To provide molecular context to our records, publicly available *Arabidopsis* resources were merged into the database, including transcriptional networks (AtRegNet from AGRIS) [25,26], *Arabidopsis* protein interactome (AI-1) [27] gene co-expression measures (ATTED-II) [28], and TAIR gene and protein annotations [24]. Thus, complex queries within KnownLeaf may combine information embedded in the Knowtator-curated annotations together with molecular data from additional public resources (Fig. 3).

3.8. Graphical representation of leaf phenotype records in a molecular context

As an alternative to structured command line queries, an interactive graph was created with the Cytoscape tool [29], in which Knowtator relations are part of a larger network of objects. The data residing in KnownLeaf were joined and simplified in several ways to produce a graph with information pertinent to interpret and further expand these relations.

Co-expression edges were only represented below a threshold value (ATTED-II co-expression mutual rank score < 25) to reduce the noise in the co-expression dataset. Next, the network was seeded with a list of AGI codes including (i) all 883 genes curated via Knowtator (Table 3) and (ii) 111 genes coding for proteins whose levels vary significantly across leaf development as determined by iTRAQ (according to more stringent cut-off criteria than previously reported in Baerenfaller, et al. [22]; global fold change > 2.8, *p*-value < 0.05) (Table S5). This combined set counted 977 non-redundant AGIs. Finally, these initial gene nodes were enriched with neighboring gene nodes through the connectivity

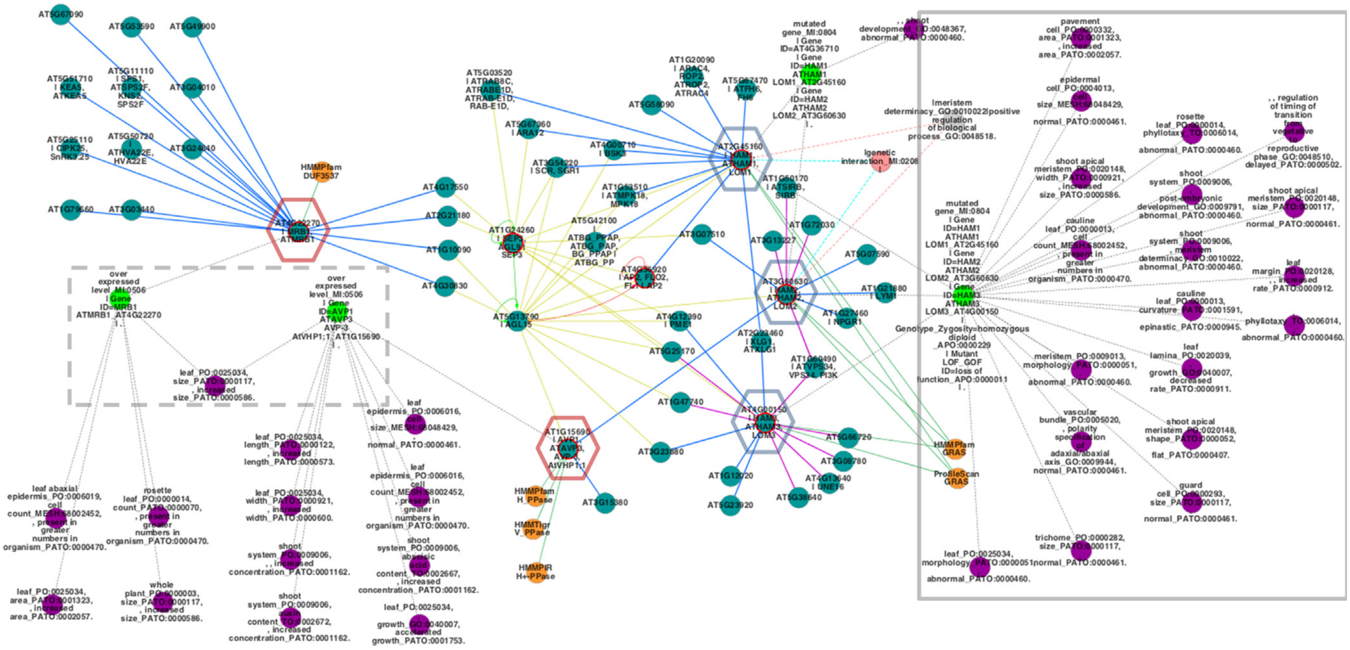

AT5G67090
AT5G53590
AT5G49900
AT5G51710 I KEA5, ATKEA5
AT5G11110 I SPS1, ATSPS2F, KNS2, SPS2F
AT3G04010
AT5G25110 I CIPK25, SnRK3.25
AT5G50720 I ATHVA22E, HVA22E
AT3G24840
HMMPfam DUF3537
AT1G79660
AT3G03440
AT4G22270 I MRB1, ATMRB1
AT4G17550
AT2G21180
AT1G10090
AT4G30830
over expressed level_MI:0506 I Gene ID=MRB1 ATMRB1_AT4G22270 I .
over expressed level_MI:0506 I Gene ID=AVP1 ATAVP3 AVP-3 AtVHP1;1_AT1G15690 I .
leaf_PO:0025034, size_PATO:0000117, , increased size_PATO:0000586.
leaf abaxial epidermis_PO:0006019, cell count_MESH:68002452, , present in greater numbers in organism_PATO:0000470.
rosette leaf_PO:0000014, count_PATO:0000070, , present in greater numbers in organism_PATO:0000470.
leaf_PO:0025034, area_PATO:0001323, , increased area_PATO:0002057.
whole plant_PO:0000003, size_PATO:0000117, , increased size_PATO:0000586.
leaf_PO:0025034, length_PATO:0000122, , increased length_PATO:0000573.
leaf_PO:0025034, width_PATO:0000921, , increased width_PATO:0000600.
shoot system_PO:0009006, , , increased concentration_PATO:0001162.
shoot system_PO:0009006, auxin content_TO:0002672, , increased concentration_PATO:0001162.
leaf epidermis_PO:0006016, cell size_MESH:68048429, , normal_PATO:0000461.
leaf epidermis_PO:0006016, cell count_MESH:68002452, , present in greater numbers in organism_PATO:0000470.
shoot system_PO:0009006, abscisic acid content_TO:0002667, , increased concentration_PATO:0001162.
leaf_PO:0025034, growth_GO:0040007, accelerated growth_PATO:0001753.
AT5G03520 I ATRAB8C, ATRABE1D, ATRAB-E1D, RAB-E1D,
AT5G67360 I ARA12
AT3G54220 I SCR, SGR1
AT4G00710 I BSK3
AT1G53510 I ATMPK18, MPK18
AT1G24260 I SEP3, AGL9 I SEP3
AT5G42100 I ATBG_PPAP, ATBG_PAP, BG_PPAP I ATBG_PP
AT4G36920 I AP2, FLO2, FL1 I AP2
AT5G13790 I AGL15
AT4G12390 I PME1
AT5G25170
AT1G47740
AT1G15690 I AVP1, ATAVP3, AVP-3, AtVHP1;1
HMMPfam H_PPase
HMMTigr V_PPase
HMMPIR H+-PPase
AT3G15380
AT3G23880
AT1G20090 I ARAC4, ROP2, ATROP2, ATRAC4
AT5G58090
AT5G67470 I ATFH6, FH6
mutated gene_MI:0804 I Gene ID=AT4G36710 I Gene ID=HAM1 ATHAM1 LOM1_AT2G45160 I Gene ID=HAM2 ATHAM2 LOM2_AT3G60630 I .
, , shoot development_GO:0048367, abnormal_PATO:0000460.
AT2G45160 I HAM1, ATHAM1, LOM1
Igenetic interaction_MI:0208 I
AT1G50170 I ATSIRB, SIRB
AT3G07510
AT3G13227
AT1G72030
AT5G07590
AT3G60630 I HAM2, ATHAM2, LOM2
AT1G21880 I LYM1
AT1G27460 I NPGR1
AT2G23460 I XLG1, ATXLG1
AT1G60490 I ATVPS34, VPS34, PI3K
AT4G00150 I HAM3, ATHAM3, LOM3
AT5G66720
AT3G08780
AT4G13640 I UNE16
AT1G12020
AT5G38640
AT5G23920
HMMPfam GRAS
ProfileScan GRAS
Imeristem determinacy_GO:0010022Ipositive regulation of biological process_GO:0048518.
mutated gene_MI:0804 I Gene ID=HAM1 ATHAM1 LOM1_AT2G45160 I Gene ID=HAM2 ATHAM2 LOM2_AT3G60630 I Gene ID=HAM3 ATHAM3 LOM3_AT4G00150 I Genotype_Zygosity=homozygous diploid _APO:0000229 I Mutant LOF_GOF ID=loss of function_APO:0000011 I .
pavement cell_PO:0000332, area_PATO:0001323, , increased area_PATO:0002057.
epidermal cell_PO:0004013, cell size_MESH:68048429, , normal_PATO:0000461.
shoot apical meristem_PO:0020148, width_PATO:0000921, , increased size_PATO:0000586.
cauline leaf_PO:0000013, cell count_MESH:68002452, , present in greater numbers in organism_PATO:0000470.
meristem_PO:0009013, morphology_PATO:0000051, , abnormal_PATO:0000460.
vascular bundle_PO:0005020, , polarity specification of adaxial/abaxial axis_GO:0009944, normal_PATO:0000461.
trichome_PO:0000282, size_PATO:0000117, , normal_PATO:0000461.
leaf_PO:0025034, morphology_PATO:0000051, , abnormal_PATO:0000460.
rosette leaf_PO:0000014, phyllotaxy_TO:0006014, , abnormal_PATO:0000460.
shoot system_PO:0009006, , post-embryonic development_GO:0009791, abnormal_PATO:0000460.
shoot system_PO:0009006, , meristem determinacy_GO:0010022, abnormal_PATO:0000460.
cauline leaf_PO:0000013, curvature_PATO:0001591, , epinastic_PATO:0000945.
leaf lamina_PO:0020039, growth_GO:0040007, decreased rate_PATO:0000911.
shoot apical meristem_PO:0020148, shape_PATO:0000052, , flat_PATO:0000407.
guard cell_PO:0000293, size_PATO:0000117, , normal_PATO:0000461.
, , regulation of timing of transition from vegetative to reproductive phase_GO:0048510, delayed_PATO:0000502.
shoot apical meristem_PO:0020148, size_PATO:0000117, , normal_PATO:0000461.
leaf margin_PO:0020128, , , increased rate_PATO:0000912.
, phyllotaxy_TO:0006014, , abnormal_PATO:0000460.

Fig. 4. LeafNet neighborhood around MRB1 and AVP1. The sub-network was derived by querying for MRB1, AVP1 and LOM* with Cytoscape's Extended Search. The color code is as follows. Nodes: AGI referring to gene or protein, teal; genotype, bright green; phenotype, purple; protein domain, orange; regulation of process, gray. Edges: AGRIS AtRegNet transcriptional regulations; yellow, red and green, representing regulation with unknown, negative and positive direction, respectively; protein–protein interactions, light purple; co-expression, blue; between AGI nodes and protein domains, light green; between AGI and genotype, orange; genetic interactions between AGIs, light blue; between AGI and phenotype, gray; between AGI and regulation of process, pink; protein–protein interaction, dark purple; indicating regulation of gene expression, olive green. Edges from external resources are solid lines, whilst Knowtator curated edges are dashed lines. The gray box highlights the nodes and edges collectively describing the phenotype of the *lom1 lom2 lom3* triple mutant. The LOM1, LOM2 and LOM3 AGI nodes are framed in blue hexagons. The dashed gray box highlights the nodes and edges indicating that *MRB1* and *AVP1* mutant plants have common leaf phenotypes. The MRB1 and AVP1 AGI nodes are framed in red hexagons.

defined by the public molecular resources incorporated within the KnownLeaf database (above).

The corresponding network, known as LeafNet, contains 19,055 nodes connected by a total of 39,649 edges, combining gene/protein–phenotype relationships with molecular information (Fig. 3). In LeafNet, the information about molecular functions, collected from the primary literature via Knowtator, complements that from the public resources. For example, 123 non-redundant protein–protein interactions were annotated involving 121 proteins. Of these, 41 (33%) overlap with those found in the AI-1 interactome, in line with the intersection commonly found between literature-curated datasets and high-throughput yeast-two hybrid datasets [7]. All the components of the software system have been made available (see Materials & Methods for details) and can be modified to create alternative network versions by adjusting threshold values or by seeding with different AGI code sets.

LeafNet is a starting point for knowledge mining. Within Cytoscape, genes, proteins or mutants can be searched with AGI codes or synonymous names (Enhanced Search plugin), their network neighborhood visualized (Select > First Neighbors of Selected Nodes), and new sub-networks created (File > New > Network > From Selected Nodes, All Edges). Combining automated and manual layout, the network context of genes/proteins of interest can be inspected and help formulate novel hypotheses. The following use-case illustrates this process with a specific example.

In *Arabidopsis*, the overexpression of both *MEMBRANE RELATED BIGGER1* (*MRB1*) and *ARABIDOPSIS VACUOLAR PYROPHOSPHATASE1* (*AVP1*) results in large leaves producing more cells, although of equal size, compared to wild type [30,31] (in red hexagons; Fig. 4). Their common phenotype is represented by connections in their LeafNet neighborhood (purple phenotype nodes and green mutant nodes in the gray box). Searching for potential regulatory relationships involving *MRB1* and *AVP1*, we noticed that they are linked in LeafNet through a path including *AGAMOUS-like 15* (*AGL15*) and *LOST MERISTEMS 2* (*LOM2*). AGL15 is a MADS-domain transcription factor that controls somatic growth [32–34]. Plants that ectopically express *AGL15* have pleiotropic mutant phenotypes, including a defect in leaf morphology [33]. LOM2 is a GRAS transcription factor [35] with two close homologs in *Arabidopsis*, LOM1 and LOM3. The *lom1 lom2 lom3* triple mutant shows an abnormal leaf morphology, indicating that all three *LOM* genes regulate cell division and cell differentiation (in dashed gray box) [36,37] (nodes in blue hexagons). This subnetwork suggests that the possible co-regulation of *MRB1* and *AVP1* by AGL15 and LOM transcription factors could be considered. While this simplified graph built with data from diverse origins may not completely or faultlessly represent actual functional links, it is useful to visualize plausible connections that warrant additional investigation.

4. Discussion

Our workflow was developed to record, among other data types, anatomical details in phenotype description (entity), what changes in that plant part or cells (attribute), and in simple terms how it changes (value). Compared to other biocuration and text mining efforts, Leaf Knowtator captures more detailed information about leaf growth and development than, for example, the more general TAIR workflow [8] or the generic large-scale text mining resource EVEX [38]. Additionally, this project was not restricted to information available in the abstract of the curated articles, but targeted any relevant sentences found in the results section of full-text articles. While the rich format of our resource resulted in a more time-consuming project, the resulting high-quality data will be a strong asset in future leaf development studies due to the high complementarity to other relevant resources.

The single most time-consuming step during annotation was the tagging of text spans in the original text of the curated article with the elements (slots) recorded in the relations. We included this constraint in the workflow to allow for semi-automated error checks, as described, and, importantly, to assist future text mining efforts. Indeed, the text mining research field is currently dominated by machine learning methods, in which lexical and grammatical patterns relevant to the problem domain are derived from manually curated training sets [39–42]. The machine learning algorithm subsequently identifies these patterns in unseen texts to predict novel annotations. The quality of such predictions relies heavily on the size and quality of the training sets. Throughout our project, we have ensured full compatibility of our methods and annotation scheme to future text mining efforts, by storing the exact offsets of the textual annotations within the original article files, thereby enabling the automatic retrieval of the specific sentence(s) and paragraph from which the information was deduced. The dimensions of our dataset are comparable to recent general-purpose annotation efforts [43], but it is unique in size and scope in the plant domain.

The specific slot-value structure defined for each relation category was also designed to facilitate future text mining efforts. By providing formal semantics and ontology terms, computer reasoning can be enhanced and well-structured annotations produced in a more time-efficient manner. For instance, the combination of manual curation, as described here, with partially automated extraction of textual information [44–46] would dramatically speed up literature curation projects. We view these opportunities as interesting follow-up work to this study.

The text corpus at the basis of this work is not exhaustive and can be expanded in several ways. Considering all primary research articles in which detailed leaf phenotypes can be linked to specific alleles of identified *Arabidopsis* genes (i.e. AGI codes), we estimate that the 283 papers we curated represent a third to a quarter of the relevant published literature. While the current dataset demonstrates the usefulness of high quality manual annotation, it would be even more valuable if it encompassed all targeted research results. Our system provides solid grounds to build up the resource by drawing in a larger community. The Leaf Knowtator interface can be adopted by any willing researcher, with documentation and training examples available to guide the first steps. Downstream software is available to monitor the consistency of the recorded relations, to transfer them into the KnownLeaf relational database, and to represent them graphically in an increased Cytoscape version of the LeafNet network.

Alternatively, Leaf Knowtator relations can be merged with large inventories of *Arabidopsis* mutations that are precious for their exhaustive coverage but that provide little detail about associated phenotypes [47,48]. Leaf Knowtator relations can also be imported into other *Arabidopsis* databases and online web tools designed to query large-scale datasets, for example TAIR [24], BAR [49], Genevestigator [11], VirtualPlant [50] or CORNET [51]. However, it is worthwhile emphasizing that such integrative systems do not yet include advanced functions to probe – beyond free text search or display – the connections between mutations in specific genes and corresponding phenotypes, stressing the usefulness of structured phenotype data enriched with ontology terms as presented here.

Through LeafNet, the integration of curated and reference knowledge resources showed that genes of interest can be placed into an informative molecular and phenotypic network landscape. LeafNet recapitulates aspects of what is already known, which is *per se* quite valuable as it joins together information dispersed in literature. In addition, it suggests new leads and close proximal associations that can be leveraged for hypothesis generation and testing. Moreover, the approach we developed could be extended and enhanced to help describe gene function in *Arabidopsis*, since a vast number of unknown or partially described genes have been placed into a molecular network landscape that goes beyond the usual GO descriptors or homology reports from sequence analysis based annotations.

An additional benefit of ontology-based phenotype descriptions is that they facilitate the comparison of phenomena between related species. For example, the Plant Ontology (PO) community is continuously improving its structured term lists to include anatomical entities that reflect the organizing principle of the plant body, thereby enabling interspecific comparisons of gene expression, phenotypes and gene functions [52]. At the other end of the research spectrum, ecologists, agronomists and breeders are also codifying trait descriptions with the implementation of dedicated ontologies to integrate field observations and measurements across experimental sites and for different species, including crops [53,54]. As our understanding of the functional modules that govern plant growth and development improves, information formatted through studies such as this one, focusing on mutant phenotypes in one plant species, could eventually assist trait development efforts in another.

To conclude, we have mustered the good will of about 15 biologists and distilled a sizable portion of the published information describing the molecular control of leaf growth and development. As the *Arabidopsis* community builds up its international bioinformatics infrastructure [55,56], we suggest that small initiatives similar to ours focusing on complementary biological domains could together contribute significantly to the inclusion of phenotype information in reference resources.

Acknowledgments

We thank Annick Bleys for help with the manuscript. We thank Dr Georgios Gkoutos for advise on the potential for machine reasoning with the curated ontological relations. This work was supported by the Integrated Project AGRON-OMICS, in the Sixth Framework Programme of the European Commission (LSHG-CT-2006-037704). S.W. was supported by TiMet - Linking the Clock to Metabolism, a Collaborative Project (Grant Agreement 245143) funded by the European Commission FP7, in response to call FP7-KBBE-2009-3. S.V.L. thanks the Research Foundation Flanders (FWO) for funding her research. A.D.J.v.D. was supported by the transPLANT project (funded by the European Commission within its 7th Framework Programme under the thematic area ‘Infrastructures’, contract number 283496). The predoctoral fellowship of J.B. was sponsored by the Agency for Innovation by Science and Technology in Flanders (IWT-Vlaanderen), a predoctoralfellowship (project no. 111164).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.cpb.2014.12.002.

References

[1] R. Hoehndorf, P.N. Schofield, G.V. Gkoutos, PhenomeNET: a whole-phenome approach to disease gene discovery, Nucleic Acids Res. 39 (2011) e119.

[2] V. Exner, P. Taranto, N. Schönrock, W. Gruissem, L. Hennig, Chromatin assembly factor CAF-1 is required for cellular differentiation during plant development, Development 133 (2006) 4163–4172.

[3] M. Krallinger, C. Rodriguez-Penagos, A. Tendulkar, A. Valencia, PLAN2L: a web tool for integrated text mining and literature-based bioentity relation extraction, Nucleic Acids Res. 37 (2009) W160–W165.

[4] A. Loyola, G. Almouzni, Histone chaperones, a supporting role in the limelight, Biochim. Biophys. Acta 1677 (2004) 3–11.

[5] H.-M. Müller, E.E. Kenny, P.W. Sternberg, Textpresso: an ontology-based information retrieval and extraction system for biological literature, PLoS Biol. 2 (2004) e309.

[6] S. Van Landeghem, S. De Bodt, Z.J. Drebert, D. Inzé, Y. Van de Peer, The potential of text mining in data integration and network biology for plant research: a case study on *Arabidopsis*, Plant Cell 25 (2013) 794–807.

[7] K. Venkatesan, J.-F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, K.-I. Goh, M.A. Yildirim, N. Simonis,

K. Heinzmann, F. Gebreab, J.M. Sahalie, S. Cevik, C. Simon, A.-S. de Smet, E. Dann, A. Smolyar, A. Vinayagam, H. Yu, D. Szeto, H. Borick, A. Dricot, N. Klitgord, R.R. Murray, C. Lin, M. Lalowski, J. Timm, K. Rau, C. Boone, P. Braun, M.E. Cusick, F.P. Roth, D.E. Hill, J. Tavernier, E.E. Wanker, A.-L. Barabási, M. Vidal, An empirical framework for binary interactome mapping, Nat. Methods 6 (2009) 83–90.

[8] D. Li, T.Z. Berardini, R.J. Muller, E. Huala, Building an efficient curation workflow for the *Arabidopsis* literature corpus, Database 2012 (2012) bas047.

[9] N. Tsesmetzis, M. Couchman, J. Higgins, A. Smith, J.H. Doonan, G.J. Seifert, E.E. Schmidt, I. Vastrik, E. Birney, G. Wu, P. D'Eustachio, L.D. Stein, R.J. Morris, M.W. Bevan, S.V. Walsh, *Arabidopsis* reactome: a foundation knowledgebase for plant systems biology, Plant Cell 20 (2008) 1426–1436.

[10] J.N. Cobb, G. DeClerck, A. Greenberg, R. Clark, S. McCouch, Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement, Theor. Appl. Genet. 126 (2013) 867–887.

[11] T. Hruz, O. Laule, G. Szabo, F. Wessendorp, S. Bleuler, L. Oertle, P. Widmayer, W. Gruissem, P. Zimmermann, Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes, Adv. Bioinform. 2008 (2008) 420747.

[12] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall, The OBI Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R.H. Scheuermann, N. Shah, P.L. Whetzel, S. Lewis, The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, Nat. Biotechnol. 25 (2007) 1251–1255.

[13] P.V. Ogren, Knowtator: a protégé plug-in for annotated corpus construction, in: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technology, Companion Volume: Demonstrations, Association for Computational Linguistics, New York City, USA, 2006, pp. 273–275.

[14] N.S.B. Miyoshi, D.G. Pinheiro, W.A. Silva Jr., J.C. Felipe, Computational framework to support integration of biomolecular and clinical data within a translational approach, BMC Bioinform. 14 (2013) 180.

[15] M. Gremse, A. Chang, I. Schomburg, A. Grote, M. Scheer, C. Ebeling, D. Schomburg, The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources, Nucleic Acids Res. 39 (2011) D507–D513.

[16] P. Jaiswal, S. Avraham, K. Ilic, E.A. Kellogg, S. McCouch, A. Pujar, L. Reiser, S.Y. Rhee, M.M. Sachs, M. Schaeffer, L. Stein, P. Stevens, L. Vincent, D. Ware, F. Zapata, Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages, Comp. Funct. Genomics 6 (2005) 388–397.

[17] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29.

[18] C. Liang, P. Jaiswal, C. Hebbard, S. Avraham, E.S. Buckler, T. Casstevens, B. Hurwitz, S. McCouch, J. Ni, A. Pujar, D. Ravenscroft, L. Ren, W. Spooner, I. Tecle, J. Thomason, C.-W. Tung, X. Wei, I. Yap, K. Youens-Clark, D. Ware, L. Stein, Gramene: a growing plant comparative genomics resource, Nucleic Acids Res. 36 (2008) D947–D953.

[19] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, R. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, C. von Mering, B. Roechert, S. Poux, E. Jung, H. Mersch, P. Kersey, M. Lappe, Y. Li, R. Zeng, D. Rana, M. Nikolski, H. Husi, C. Brun, K. Shanker, S.G.N. Grant, C. Sander, P. Bork, W. Zhu, A. Pandey, A. Brazma, B. Jacq, M. Vidal, D. Sherman, P. Legrain, G. Cesareni, L. Xenarios, D. Eisenberg, B. Steipe, C. Hogue, R. Apweiler, The HUPO PSI's molecular interaction format – a community standard for the representation of protein interaction data, Nat. Biotechnol. 22 (2004) 177–183.

[20] J. Hannemann, H. Poorter, B. Usadel, O.E. Bläsing, A. Finck, F. Tardieu, O.K. Atkin, T. Pons, M. Stitt, Y. Gibon, Xeml Lab: a tool that supports the design of experiments at a graphical interface and generates computer-readable metadata files, which capture information about genotypes, growth conditions, environmental perturbations and sampling strategy, Plant Cell Environ. 32 (2009) 1185–1200.

[21] H. Poorter, F. Fiorani, M. Stitt, U. Schurr, A. Finck, Y. Gibon, B. Usadel, R. Munns, O.K. Atkin, F. Tardieu, T.L. Ponsi, The art of growing plants for experimental purposes: a practical guide for the plant biologist, Funct. Plant Biol. 39 (2012) 821–838.

[22] K. Baerenfaller, C. Massonnet, S. Walsh, S. Baginsky, P. Bühlmann, L. Hennig, M. Hirsch-Hoffmann, K.A. Howell, S. Kahlau, A. Radziejwoski, D. Russenberger, D. Rutishauser, I. Small, D. Stekhoven, R. Sulpice, J. Svozil, N. Wuyts, M. Stitt, P. Hilson, C. Granier, W. Gruissem, Systems-based analysis of *Arabidopsis* leaf growth reveals adaptation to water deficit, Mol. Syst. Biol. 8 (2012) 606.

[23] C. Massonnet, D. Vile, J. Fabre, M.A. Hannah, C. Caldana, J. Lisec, G.T.S. Beemster, R.C. Meyer, G. Messerli, J.T. Gronlund, J. Perkovic, E. Wigmore, S. May, M.W. Bevan, C. Meyer, S. Rubio-Díaz, D. Weigel, J.L. Micol, V. Buchanan-Wollaston, F. Fiorani, S. Walsh, B. Rinn, W. Gruissem, P. Hilson, L. Hennig, L. Willmitzer, C. Granier, Probing the reproducibility of leaf growth and molecular phenotypes: a comparison of three *Arabidopsis* accessions cultivated in ten laboratories, Plant Physiol. 152 (2010) 2142–2157.

[24] P. Lamesch, T.Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D.L. Alexander, M. Garcia-Hernandez, A.S. Karthikeyan, C.H. Lee, W.D. Nelson, L. Ploetz, S. Singh, A. Wensel, E. Huala, The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools, Nucleic Acids Res. 40 (2012) D1202–D1210.

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

[26] A. Yilmaz, M.K. Mejia-Guerra, K. Kurz, X. Liang, L. Welch, E. Grotewold, AGRIS: the *Arabidopsis* Gene Regulatory Information Server, an update, Nucleic Acids Res. 39 (2011) D1118–D1122.

[27] Arabidopsis Interactome Mapping Consortium, Evidence for network evolution in an *Arabidopsis* interactome map, Science 333 (2011) 601–607.

[28] T. Obayashi, S. Hayashi, M. Saeki, H. Ohta, K. Kinoshita, ATTED-II provides coexpressed gene networks for *Arabidopsis*, Nucleic Acids Res. 37 (2009) D987–D991.

[29] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504.

[30] N. Gonzalez, S. De Bodt, R. Sulpice, Y. Jikumaru, E. Chae, S. Dhondt, T. Van Daele, L. De Milde, D. Weigel, Y. Kamiya, M. Stitt, G.T.S. Beemster, D. Inzé, Increased leaf size: different means to an end, Plant Physiol. 153 (2010) 1261–1279.

[31] H. Guan, D. Kang, M. Fan, Z. Chen, L.-J. Qu, Overexpression of a new putative membrane protein gene *AtMRB1* results in organ size enlargement in *Arabidopsis*, J. Integr. Plant Biol. 51 (2009) 130–139.

[32] B.J. Adamczyk, M.D. Lehti-Shiu, D.E. Fernandez, The MADS domain factors AGL15 and AGL18 act redundantly as repressors of the floral transition in *Arabidopsis*, Plant J. 50 (2007) 1007–1019.

[33] D.E. Fernandez, G.R. Heck, S.E. Perry, S.E. Patterson, A.B. Bleecker, S.-C. Fang, The embryo MADS domain factor AGL15 acts postembryonically: inhibition of perianth senescence and abscission via constitutive expression, Plant Cell 12 (2000) 183–197.

[34] E.W. Harding, W. Tang, K.W. Nichols, D.E. Fernandez, S.E. Perry, Expression and maintenance of embryogenic potential is enhanced through constitutive expression of *AGAMOUS-Like 15*, Plant Physiol. 133 (2003) 653–663.

[35] C. Bolle, The role of GRAS proteins in plant signal transduction and development, Planta 218 (2004) 683–692.

[36] E.M. Engstrom, C.M. Andersen, J. Gumulak-Smith, J. Hu, E. Orlova, R. Sozzani, J.L. Bowman, *Arabidopsis* homologs of the *Petunia HAIRY MERISTEM* gene are required for maintenance of shoot and root indeterminacy, Plant Physiol. 155 (2011) 735–750.

[37] S. Schulze, B.N. Schäfer, E.A. Parizotto, O. Voinnet, K. Theres, LOST MERISTEMS genes regulate cell differentiation of central zone descendants in *Arabidopsis* shoot meristems, Plant J. 64 (2010) 668–678.

[38] S. Van Landeghem, J. Björne, C.-H. Wei, K. Hakala, S. Pyysalo, S. Ananiadou, H.-Y. Kao, Z. Lu, T. Salakoski, Y. Van de Peer, F. Ginter, Large-scale event extraction from literature with multi-level gene normalization, PLOS ONE 8 (2013) e55814.

[39] J. Björne, F. Ginter, T. Salakoski, University of Turku in the BioNLP'11 Shared Task, BMC Bioinform. 13 (2012) S4.

[40] D. McClosky, S. Riedel, M. Surdeanu, A. McCallum, C.D. Manning, Combining joint models for biomedical event extraction, BMC Bioinform. 13 (2012) S9.

[41] R. Sætre, K. Yoshida, M. Miwa, T. Matsuzaki, Y. Kano, J. Tsujii, Extracting protein interactions from text with the unified AkaneRE event extraction system, IEEE-ACM Trans. Comput. Biol. Bioinform. 7 (2010) 442–453.

[42] Y. Zhang, H. Lin, Z. Yang, J. Wang, Y. Li, A single kernel-based approach to extract drug–drug interactions from biomedical literature, PLOS ONE 7 (2012) e48901.

[43] C. Nédellec, R. Bossy, J.-D. Kim, J.-J Kim, T. Ohta, S. Pyysalo, P. Zweigenbaum, Overview of BioNLP Shared Task 2013, in: Proceedings of the BioNLP Workshop, Sofia, Bulgaria, 9 August 2013, 2013, pp. 1–7.

[44] C.N. Arighi, P.M. Roberts, S. Agarwal, S. Bhattacharya, G. Cesareni, A. Chatr-Aryamontri, S. Clematide, P. Gaudet, M.G. Giglio, I. Harrow, E. Huala, M. Krallinger, U. Leser, D. Li, F. Liu, Z. Lu, L.J. Maltais, N. Okazaki, L. Perfetto, F. Rinaldi, R. Sætre, D. Salgado, P. Srinivasan, P.E. Thomas, L. Toldo, L. Hirschman, C.H. Wu, BioCreative III interactive task: an overview, BMC Bioinform. 12 (2011) S4.

[45] L. Hirschman, G.A.P.C. Burns, M. Krallinger, C. Arighi, K.B. Cohen, A. Valencia, C.H. Wu, A. Chatr-Aryamontri, K.G. Dowell, E. Huala, A. Lourenço, R. Nash, A.-L. Veuthey, T. Wiegers, A.G. Winter, Text mining for the biocuration workflow, Database 2012 (2012) bas020.

[46] C.-H. Wei, H.-Y. Kao, Z. Lu, PubTator: a web-based text mining tool for assisting biocuration, Nucleic Acids Res. 41 (2013) W518–W522.

[47] J. Lloyd, D. Meinke, A comprehensive dataset of genes with a loss-of-function mutant phenotype in *Arabidopsis*, Plant Physiol. 158 (2012) 1115–1129.

[48] D.W. Meinke, A survey of dominant mutations in *Arabidopsis thaliana*, Trends Plant Sci. 18 (2013) 84–91.

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

[50] K. Toufighi, S.M. Brady, R. Austin, E. Ly, N.J. Provart, The botany array resource: e-Northerns, expression angling, and promoter analyses, Plant J. 43 (2005) 153–163.

[51] S. De Bodt, J. Hollunder, H. Nelissen, N. Meulemeester, D. Inzé, CORNET 2.0: integrating plant coexpression, protein-protein interactions, regulatory interactions, gene associations and functional annotations, New Phytol. 195 (2012) 707–720.

[52] L. Cooper, R.L. Walls, J. Elser, M.A. Gandolfo, D.W. Stevenson, B. Smith, J. Preece, B. Athreya, C.J. Mungall, S. Rensing, M. Hiss, D. Lang, R. Reski, T.Z. Berardini, D. Li, E. Huala, M. Schaeffer, N. Menda, E. Arnaud, R. Shrestha, Y. Yamazaki, P. Jaiswal, The Plant Ontology as a tool for comparative plant anatomy and genomic analyses, Plant Cell Physiol. 54 (2013) e1 (1–23).

[53] R. Shrestha, L. Matteis, M. Skofic, A. Portugal, G. McLaren, G. Hyman, E. Arnaud, 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 (2012) 326.

[54] N. Pérez-Harguindeguy, S. Díaz, E. Garnier, S. Lavorel, H. Poorter, P. Jaureguiberry, M.S. Bret-Harte, W.K. Cornwell, J.M. Craine, D.E. Gurvich, C. Urcelay, E.J. Veneklaas, P.B. Reich, L. Poorter, I.J. Wright, P. Ray, L. Enrico, J.G. Pausas, A.C. de Vos, N. Buchmann, G. Funes, F. Quétier, J.G. Hodgson, K. Thompson, H.D. Morgan, H. ter Steege, M.G.A. van der Heijden, L. Sack, B. Blonder, P. Poschlod, M.V. Vaieretti, G. Conti, A.C. Staver, S. Aquino, J.H.C. Cornelissen, New handbook for standardised measurement of plant functional traits worldwide, Aust. J. Bot. 61 (2013) 167–234.

[55] International Arabidopsis Informatics Consortium, An international bioinformatics infrastructure to underpin the *Arabidopsis* community, Plant Cell 22 (2010) 2530–2536.

[56] The International Arabidopsis Informatics Consortium, Taking the next step: building an *Arabidopsis* Information Portal, Plant Cell 24 (2012) 2248–2256.

[57] D. Fleury, K. Himanen, G. Cnops, H. Nelissen, T.M. Boccardi, S. Maere, G.T. Beemster, P. Neyt, S. Anami, P. Robles, J.L. Micol, D. Inzé, M. Van Lijsebettens, The Arabidopsis thaliana homolog of yeast BRE1 has a function in cell cycle regulation during early leaf and root growth, Plant Cell 19 (2) (2007) 417–432.