It's All GO for Plant Scientists[1]

Jennifer I. Clark*, Cath Brooksbank, and Jane Lomax

EMBL-EBI Wellcome Trust Genome Campus Hinxton, Cambridge CB10 1SD, United Kingdom

The Gene Ontology project (http://www. geneontology.org/) produces structured, controlled vocabularies and gene product annotations. Gene products are classified according to the cellular locations and biological process in which they act, and the molecular functions that they carry out. We annotate gene products from a broad range of model species and provide support for those groups that wish to contribute annotation of further model species. The Gene Ontology facilitates the exchange of information between groups of scientists studying similar processes in different model organisms, and so provides a broad range of opportunities for plant scientists.

THE GOAL OF GENE ONTOLOGY

Rapid innovation in biology has given rise to vast amounts of biological data. Biologists wish to draw on related research carried out on different model species but are currently hampered by differences in technical language. For example, a plant scientist might try to access information on gene products involved in the process of gametogenesis, the development of the plant gametophyte (in doing this they might read, for example, Robertson et al., 2004). During their search for other relevant publications, they would be frustrated to find papers describing gene products involved in another kind of gametogenesis (e.g. Ma et al., 2004). This is because the same word is used in animal biology to describe the process of production of gametes in animals. It is difficult and frustrating for a scientist to sort out the meaning of biological language when the same undefined word is used for two different concepts, and this task is impossible for a computer.

The aim of the Gene Ontology (GO) project (Ashburner et al., 2000; Harris et al., 2004) is to provide a standard language for the description of gene products. To address this aim, the GO project is developing ontologies and using them in annotation of gene products. We aim to support annotation of gene products from as many species as possible, so that for any specific research topic, information can be easily transferred between groups studying a diversity of model species. (See the GO Web site for a full list of member organizations, http://www.geneontology.org/GO.consortiumlist.html.) We are also setting up tools and systems to provide biologists with free and open access to the data generated.

The ontologies are catalogs containing the full range of biological processes, molecular functions, and cellular components needed to describe all gene products. As an example, the gene product cytochrome c can be described as having the molecular function "electron transporter activity," as being involved in the biological processes "oxidative phosphorylation" and "induction of cell death," and as acting in the location of the cellular components "mitochondrial matrix" and "mitochondrial inner membrane." Creating and recording descriptions of gene products in terms of the concepts captured in GO is known as annotation of gene products to GO.

The controlled vocabularies are structured so that the user can query them at different levels. For example, a user can search the GO (http://www.godatabase.org/) for gene products involved in the process of gametogenesis and be automatically offered lists of gene products involved in gametophyte development. These gene product listings provide links to the primary research papers showing experimental evidence for the gene product's involvement in the given process. The GO includes both general and very specific concepts and a broad range of species. Users can find all the gene products that are involved in a very general process like signal transduction or zoom in on a specific process like ethylene biosynthesis.

In addition to simplifying the process of searching the scientific literature, the GO facilitates analysis of experimental results. For example, use of the ontologies and annotations in combination simplifies the interpretation of microarray data. Using GO, scientists can extract meaningful conclusions from data on the expression patterns of thousands of genes to examine the effect of treatment on particular processes. The GO can be used to identify which biological processes are overrepresented among those genes whose expression is altered. This method has been used in various studies including those on the effect of temperature reduction (Malek et al., 2004) and desiccation (Oliver et al., 2004) on gene expression profiles. In addition, the GO has been used to analyze gene function in studies of the evolution of duplicated genes (Blanc and

[1] This work was supported by the National Institutes of Health/National Human Genome Research Institute (grant no. HG02273) and by the European Union Research and Technological Development Quality of Life and Management of Living Resources Programme (QLRI–CT–2001–00981 and QLRI–CT–2001–00015).

* Corresponding author; e-mail jclark@ebi.ac.uk; fax 44–(0)1223–494–468.

www.plantphysiol.org/cgi/doi/10.1104/pp.104.058529.

Wolfe, 2004), alternative splicing (Zhou et al., 2003), and fungal development and pathogenesis (Li et al., 2004). A further use of the GO is to give an overview of the genome of newly sequenced organisms (Goff et al., 2002; Berardini et al., 2004). This gives an accurate snapshot of the range of gene products encoded in the genome of a newly sequenced species. More examples of the use of GO can be found in the GO bibliography (http://www.geneontology.org/GO.biblio.html).

As use of the GO becomes more widespread, there is an increasing need to provide a concise introduction to the GO project's resources and practices. The purpose of this paper is to give a clear explanation of the ontologies and annotations in the context of biological research. We seek to accommodate primarily those who simply wish to understand the GO and search the information online, but we also provide an introduction for those who may go on to provide gene product annotations or to make more extensive computational use of the data. We describe first the content, structure, and development of the ontologies, and then consider how the ontologies are applied in gene product annotation, emphasizing the issues and opportunities facing plant biologists. The last part of the paper covers very practical issues including software tools, file formats, and GO slims.

HOW ARE THE ONTOLOGIES STRUCTURED?

The ontologies resemble dictionaries. As in a dictionary, concepts are captured as terms and have a term name (word or phrase) and a text definition. For example, the definition "Interacting selectively with DNA (deoxyribonucleic acid)" is given for the concept with the name "DNA binding." In GO, if one concept is known by many different names, then a single term name would be entered and the other names would be entered as synonyms. For example, the concept tricarboxylic acid cycle would have a single term name with the synonyms Krebs cycle, TCA cycle, and citric acid cycle. Conversely, where two groups of scientists use a single word or phrase in different ways, two GO terms will be created, with the term names distinguished by the addition of a sensu designation. The sensu designation shows that a word is used in a given "sense." For example, in the case of the term gametogenesis, the plant-specific term name is "gametogenesis (sensu Magnoliophyta)" (gametogenesis in the sense used in flowering plant biology), while the term describing the concept as used by animal biologists is named "gametogenesis (sensu Metazoa)" (gametogenesis in the sense used in animal biology).

GO extends the idea of a dictionary, so that in GO if one concept could be considered to be related to

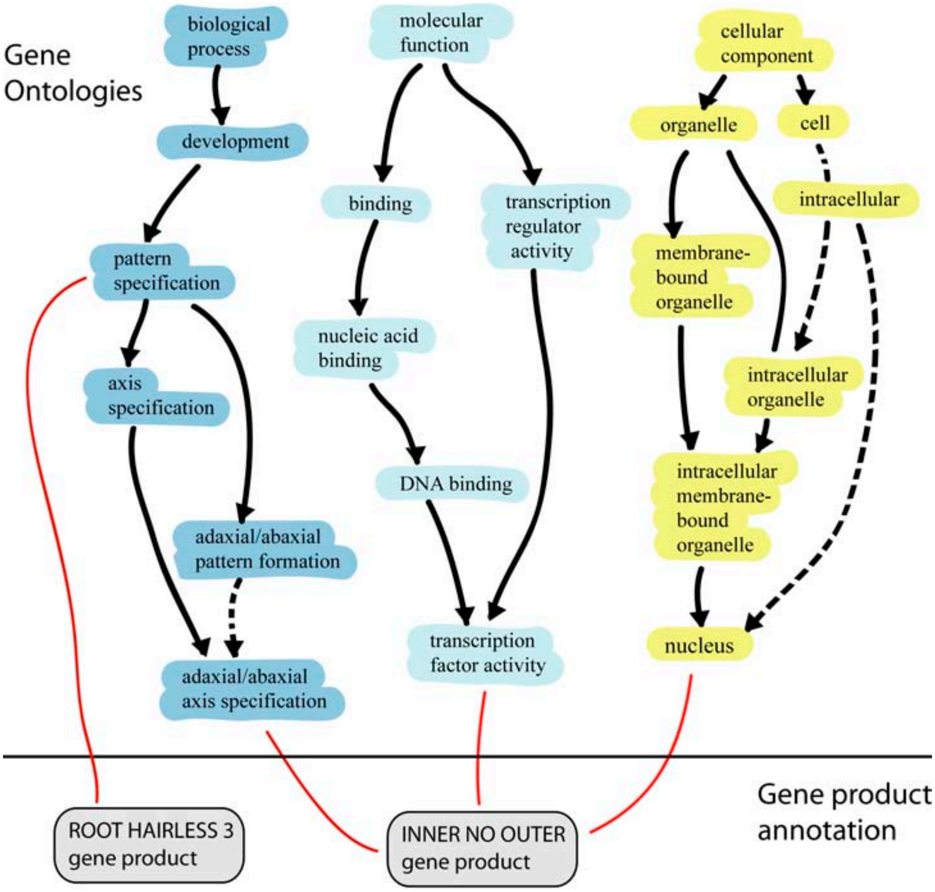


Figure 1. The ontologies. Here, sections of the three ontologies are represented schematically with only term names shown. The biological process ontology is shown on the left side (dark blue background), the molecular function ontology is shown in the center (light blue background), and the cellular component ontology is shown on the right side (yellow background). More general concepts are at the top and the more specific ones are at the bottom. The is_a relationships (continuous black lines) indicate that a child concept is a type of the parent concept, and the part_of relationships (dashed black lines) indicate that the child concept is a part of the parent concept. A term may have multiple parent terms, as in the case of adaxial/abaxial axis specification. Separately and in parallel with the development of the GO, gene products are annotated (red lines) to the terms. Annotation indicates that the gene product (gray background, black outline) is involved in the process described, or that it has the function, or acts in the location described.

another concept, a link is made between the two (depicted in Fig. 1). This is in contrast to the simple alphabetical listing of concepts in a dictionary, where no comprehensive attempt is made to link terms to one another. A concept in GO can be a type of (or have the ''is_a'' relationship to) another concept and can also be a part of (have the ''part_of'' relationship to) another concept. The GO is able to represent multiple parentage through use of an arrangement known as a directed acyclic graph, which resembles a hierarchy, but differs in that the structure allows a child (more specialized term) to have many parents (less specialized terms). Furthermore, every GO term must inherit all the properties of any parent terms. In terms of gene product annotation, if the child term describes the gene product, then all its parent terms must also apply to that gene product. Finally, each GO term is assigned a unique numerical identifier.

CONTENT OF THE ONTOLOGIES

The GO project provides three separate ontologies to capture molecular functions, biological processes, and cellular components. GO molecular function terms represent the activities of gene products. They are generally single step reactions, such as alcohol dehydrogenase activity that describes the catalysis of the reaction: an alcohol + NAD^+ = an aldehyde or ketone + NADH + H^+. Further examples of molecular functions include ''transcription factor activity'' and ''ribulose-bisphosphate carboxylase activity.'' For information on cross-references between the GO and the Enzyme Commission Enzyme Nomenclature, see the section titled Cross-references. GO biological process terms describe ordered assemblies of molecular functions. These include broad terms like respiration and more specific terms like PSII assembly and primary charge separation. The GO cellular component ontology includes any component of a cell, and this may be an anatomical structure (for example chloroplast, secondary cell wall, or nucleus) or a gene product group (for example PSII associated light-harvesting complex II, peripheral complex). For more information on the content of the three ontologies, see the GO documentation available online at http://www.geneontology.org/GO.contents.doc.html.

UPDATING THE ONTOLOGIES

GO is a work in progress, so the ontologies should not be considered to be complete or static. The Consortium consists of a wide range of groups that cooperate to arrive at a consensus. In response to discussions within the community, the ontologies are constantly updated and refined, with new terms being added daily to accommodate the needs of annotators. The ontologies are edited using the tool, OBO-Edit (discussed later). Table I shows the number of terms in GO over time.

There are many reasons to update the ontologies. Each new GO term is now carefully defined when it is added. However, some of the earliest terms were not defined when they were initially added and these terms are still gradually being defined. The great progress that has been made in retrospectively applying these definitions can be seen in Tables I and II. The research community can recommend changes to the GO, such as making definitions broader or more

Table I. *Changes in the ontologies*

The Gene Ontologies are under continuous development and changes are made on a daily basis. Changes in the number of defined, undefined, and obsolete terms since January 2001 are shown.

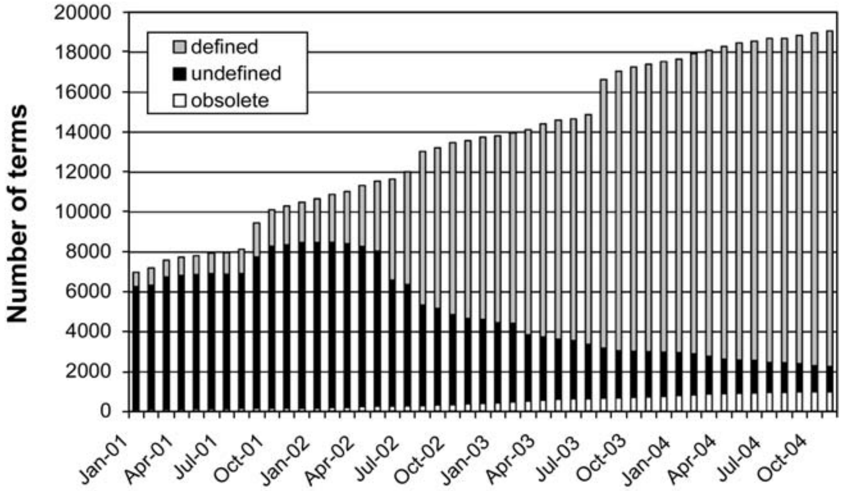


Table II. *Numbers of terms in each ontology, as calculated using the OBO format ontology file of December 16, 2004*

	Biological Process	Molecular Function	Cellular Component
Terms	9,235	7,455	1,510
Obsolete terms	4%	7%	7%
Nonobsolete defined terms	96%	90%	94%

specific. If the new definition is just a change in wording and the term still describes the same concept, then it will still be kept as a working term and it will retain its identification number. However, if a term needs to be radically redefined so that it no longer describes the same concept as before and so that annotations to the term would no longer be correct, then the old term will become obsolete and a new term created with a new identification number. Obsoletion means that the term will still be present within the ontology but will no longer be used for annotation and will be moved from its usual place in the ontology and placed under an obsolete node. The obsolete terms may be viewed by expanding the obsolete nodes visible in the front page of the AmiGO browser at http://www.godatabase.org/. If a term has been used in annotation and it then becomes obsolete then annotations will be transferred to another more suitable term by the curators in the database maintaining the annotations. Reallocating annotation to a new replacement term is very time consuming for curators, so we always try to avoid obsoletion if at all possible. An obsoletion will only go ahead once it is agreed by all interested parties. Table II shows the current total number of terms, the percentage of obsolete terms, and the percentage of defined nonobsolete terms.

In addition to improving terms already present, curators often need to request new terms. For example, a curator may wish to annotate the gene products involved in the process of petal lobe development and find that only the term petal development is present. In this case, the curator could request the addition of the new term and suggest possible parentage and a definition. This suggestion would be entered in the curator request tracker on the GO Web discussion forum (details below), where it would be taken up by a curator at the editorial office. In processing a request, the curator checks that no appropriate term is present and then discusses with all the relevant groups whether the term should be added, where it should go, what it should be called, and how it should be defined. Literature references are used and are cited as a "dbxref." Once the term is agreed, the curator adds it, to be available for use in annotation the following day. New terms are shown in the monthly report and in the daily e-mail of changes.

Another situation that sometimes arises is that a new species is being annotated and, though the relevant process is already in GO, the term reflects the process as it is seen in one group of organisms and does not adequately reflect the nature of the process in the new species. In this case, the consortium discusses ways to alter the relevant parts of the ontology to accommodate the differences between species, and once agreement is reached the new system is implemented. An example currently in progress is an attempt to modify the terms describing flower development in Arabidopsis (*Arabidopsis thaliana*) so that they will accommodate the development of flowers in grass species.

The Consortium actively encourages the involvement of biologists with expertise in relevant biological domains. We need expert input on how the GO can accurately reflect the language used within the scientific community and the current state of scientific knowledge. There are various ways by which experts can become involved. Often the curators search for somebody in the given subject area who is geographically close to them. Alternatively, the experts may come forward and volunteer their services, as happened recently with the PAMGO (plant-associated microbe gene ontology) group. Where a group like this is willing to contribute, a curator is assigned to guide them through the process of developing the ontology structure. The curator and experts work together to solve the ontology problem and develop the new terms. If the topic is very large or if many different views need to be taken into account, then a meeting can be held where the various different view points can be represented and conclusions can be reached that will accommodate everyone's requirements. In the case of the PAMGO group, the experts were guided through the complexities of ontology development by a curator at The Institute for Genomic Research (TIGR), and a meeting was hosted by The Arabidopsis Information Resource (TAIR). Through this process, terms to cover pathogenesis were developed. The terms that were agreed upon at the meeting were posted for comment on the Web discussion forum and then presented at a consortium meeting. The minutes for both of these meetings are available on the GO Web site. Finally, the terms were added to the process ontology by a curator at the editorial office and are now available for use. By collaborations such as this, the consortium seeks to incorporate the full range of processes needed for annotation of gene products, including all plant gene products. To see the terms that were added, view the term "interaction between organisms" and its children in AmiGO. This kind of collaboration is one way in which terms can come to be added or improved, and we would welcome further approaches by groups with specific interest in given biological domains.

The vast majority of our discussions occur online and can be viewed in the Web discussion forum at http://geneontology.sourceforge.net/. These discussions are supplemented by e-mail that is archived at http://www.geneontology.org/go_email.html, and by Consortium meetings, the minutes of which are available online at http://www.geneontology.org/GO.meetings.html.

Since the ontologies are constantly under development it is important that users who download a copy of the GO for analysis note, and include in publications, the release number, or the date and format of the version that they use. For those users who prefer a monthly release the ontologies are archived at ftp://ftp.geneontology.org/pub/go/ontology-archive/.

METHOD OF ANNOTATION

Annotation has been explained, above, as a system whereby gene products can be described using the GO controlled vocabularies. Figure 2 shows a schematic diagram of gene products, derived from a range of species, which have been annotated to a single GO term. Annotation of several different gene products to the same GO term indicates that the gene products have the same function or act in the same cellular location or biological process. All of the gene products annotated in Figure 2 have calcium-transporting ATPase activity. The individual gene products in this diagram were annotated by a range of database resources as indicated, and then all the annotation information from each database resource was contributed to the central repository at the GO Consortium. Collection of annotation data in this way allows users to search across the full dataset and range of species using a single query. The collection of data relies on the collaboration of the large number of bioinformatics resources in which annotation is carried out. Annotation data is primarily produced in species-specific database resources, such as the Saccharomyces Genome Database (SGD), the Mouse Genome Informatics (MGI) resource and FlyBase, and in multispecies resources such as UniProt. For plant gene products, the main contributors are currently TIGR, TAIR, Gramene, and UniProt. For a complete list of contributing database groups and for the total numbers of annotations, see http://www.geneontology.org/GO.current.annotations.shtml.

Annotations may be made computationally (sometimes called electronic annotation) or curatorially (manually by an annotator/curator). Manual annotation methods involve a curator reading the relevant primary research papers and extracting and recording information. Annotation requires detailed knowledge of the relevant biological domain, rather than extensive bioinformatics experience, and so is best done by the experimental biologists themselves. For the more heavily used model species, the curators will be biologists who have converted to be full-time annotators as staff of a large database resource. In cases where a model species is not yet completely sequenced, or where there is not yet an established database resource, curators may also be full-time bench scientists.

To make an annotation from a paper, there are four crucial pieces of information that the annotator must find. The first is the accession number for the gene product being described, and the second is the identification number for the GO term describing the gene product. The third piece of information is the reference number of the paper or information source in which the evidence was found, and the last is the evidence code indicating what kind of evidence supports the assertion that the GO term describes the gene product. Additional information can also be captured; to view the list, see Table III and http://www.geneontology.org/GO.annotation.shtml.

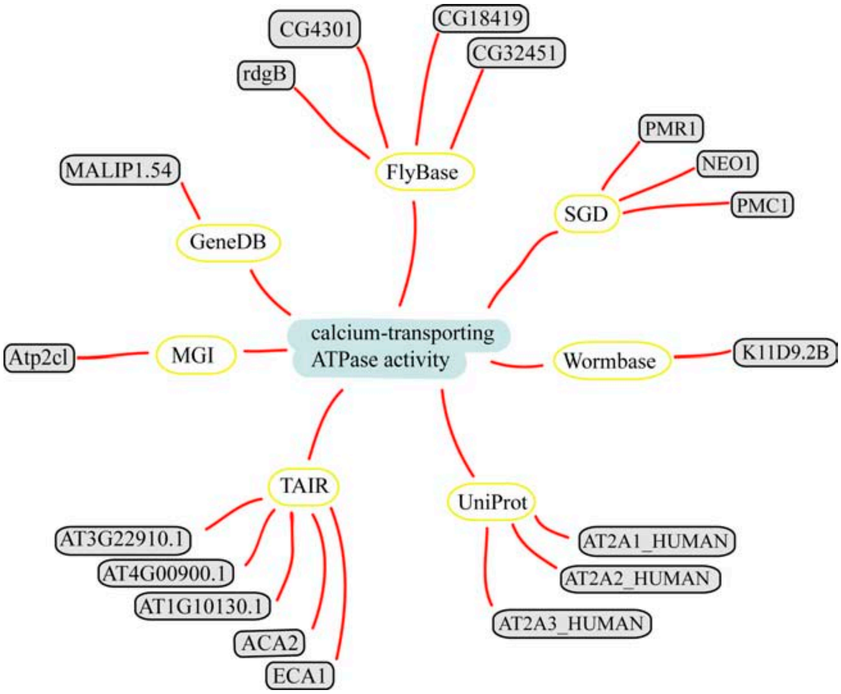


Figure 2. Annotation. A wide variety of gene products (gray background, black outline) may be annotated (red lines) to a single GO function term (light blue background). The annotations are made by curators in a range of bioinformatics database resources, for example MGI, TAIR, and Wormbase (text in yellow circles). The annotation of multiple gene products to a single GO function term (as shown here) indicates that they mediate the same single step reaction. The database resource acronyms are expanded as follows: Saccharomyces Genome Database (SGD), UniProt Knowledgebase (UniProt), The Arabidopsis Information Resource (TAIR), Mouse Genome Informatics (MGI), and *Plasmodium falciparum* GeneDB (GeneDB).

Table III. *Gene association file column contents*

The fifteen columns of the gene association file each contain a specific type of information, as described here. Each piece of information documents some aspect of the connection of the gene product with a GO term, or the evidence that supports that association.

Column	Content
1	Acronym of the database contributing the annotation
2	Identifier for the gene product annotated, as used within the database contributing the annotation
3	Symbol for the gene product
4	Flags that modify the interpretation of an annotation (see GO Web site)
5	GO numerical identifier for the term to which the gene product is being annotated
6	Reference no. for the publication on which the annotation is based
7	Evidence code (see Table IV)
8	Additional identifier (see the GO Web site)
9	Ontology: P (biological process), F (molecular function), or C (cellular component)
10	Name of gene or gene product
11	Gene_symbol
12	Shows what kind of thing is annotated: gene, transcript, protein, protein_structure, or complex
13	Taxonomic identifier for the species encoding the gene product
14	Date on which the annotation was made
15	The database that made the annotation

The process of finding the information for an annotation can be demonstrated by the example of annotation of the protein Ser/Thr kinase function of the PERK1 gene product of *Brassica napus*. An annotator may find a paper reporting biochemical experiments showing that the kinase domain of PERK1 has Ser/Thr kinase activity (Silva and Goring, 2002). To make an annotation, the annotator would search for the four crucial pieces of information described above. The accession number of the gene product can easily be found from various online bioinformatics resources. The annotator would search the GO to find the most appropriate term to describe the activity of the gene product (perhaps using the tool AmiGO, described later). In this case, it would be "protein Ser/Thr kinase activity" (GO:0004674). The annotator would note the reference number for the paper, and then the evidence code showing the nature of the supporting evidence. The level of reliability of an annotation depends on the evidence on which the annotation was based. The different types of evidence that may be presented have been categorized by the GO Consortium to simplify presentation and interpretation of the information, and each has an evidence code. For example, TAS stands for traceable author statement and IDA for inferred from direct assay (Table IV shows all evidence codes currently in use and the types of evidence that they represent). In the case of this annotation of the PERK1 protein, an experimental assay showed the function of the gene product and so the annotation will carry the evidence code IDA, indicating that the conclusion in the paper about the gene function was inferred from direct assay. Multiple annotations may be made for a single gene product. For example, the paper about PERK1 may state further information about other functions of the gene product or the location or processes in which the gene product acts. If this is the case, then the gene product could be annotated to multiple terms from all three ontologies to capture all of the relevant information. All the information constituting a gene product annotation is noted down in a gene association file, which may contain a large number of annotations, each taking up one line of the file. The file is then sent off to the GO Consortium for storage and redistribution. Each annotating group within the consortium submits annotation data to the central GO repository in a tab-delimited file format called the gene association file (shown in Fig. 3) and the GO Consortium makes the data available online. For further information on the file format, see http://www.geneontology.org/GO.annotation.html#file.

The example of PERK1 annotation illustrates the manual annotation of gene products, but annotation may also be made computationally (electronic annotation). Electronic methods involve extrapolation of annotations by comparison of unannotated gene products with those that have been extensively manually annotated. A number of different automatic methods have been applied (Pouliot et al., 2001; Okazaki et al., 2002; Xie et al., 2002; Camon et al., 2003; Mi et al., 2003), and third-party online tools that can generate electronic annotations are available (e.g. InterProScan; Zdobnov and Apweiler, 2001). Electronic annotations generated by any of these methods are represented by the evidence code IEA (inferred from electronic annotation). In addition to the small set of extensively manually annotated model organisms, the electronic annotations of most species can be found in the UniProt database (http://www.uniprot.org).

There are important differences between manual and electronic annotation. Curators of biological literature are generally biologists who are knowledgeable about an experimental organism and about molecular biology. They produce very high quality annotation that draws exclusively on published experimental results, and this is time-consuming. Electronic methods, conversely, produce huge numbers of lower quality annotations very quickly. These electronic methods are especially useful for the annotation of gene products that are less amenable to experimental methods, such as those of humans. The quality of electronic annotation has recently been assessed in some detail (Camon et al., 2005). This research found that in the worst case scenario, the generation

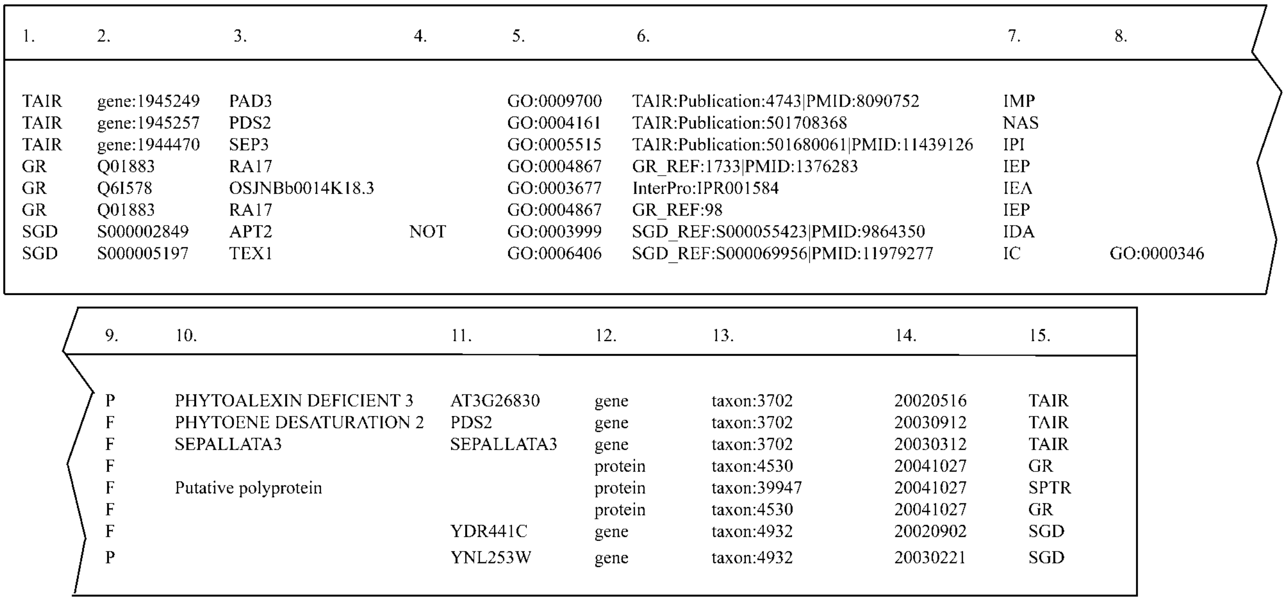
Table IV. *Evidence codes*

The evidence codes show the nature of the evidence supporting the assertion that a gene product is described by a given process, function, or component term. The evidence codes are listed here with their acronyms, expanded names, and some examples of the types of studies each code represents.

Evidence Code Acronym	Full Name of Evidence Code	Examples
IDA	Inferred from direct assay	Physical interaction/binding Enzyme assays In vitro reconstitution (e.g. transcription) Immunofluorescence (for cellular component) Cell fractionation (for cellular component) Physical interaction/binding
IEP	Inferred from expression pattern	Transcript levels (e.g. northerns, microarray data) Protein levels (e.g. western blots)
IGI	Inferred from genetic interaction	"Traditional" genetic interactions such as suppressors, synthetic lethals, etc. Functional complementation Rescue experiments Inference about one gene drawn from the phenotype of a mutation in a different gene
IMP	Inferred from mutant phenotype	Any gene mutation/knockout Overexpression/ectopic expression of wild-type or mutant genes Antisense experiments RNAi experiments Specific protein inhibitors
IPI	Inferred from physical interaction	2-Hybrid interactions Copurification Coimmunoprecipitation Ion/protein binding experiments
ISS	Inferred from sequence similarity	Sequence similarity or structural similarity (homolog of/most closely related to) Recognized domains Structural similarity Southern blotting
RCA	Inferred from reviewed computational analysis	Predictions based on large-scale protein interaction experiments Predictions based on microarray results Predictions based on integration of large-scale data sets of several types Text-based computation (e.g. text mining)
IEA	Inferred from electronic annotation	
IC	Inferred by curator	
NAS	Nontraceable author statement	
TAS	Traceable author statement	
ND	No biological data available	

of electronic annotations using the interpro2go, spkw2go, and ec2go mapping files precisely predicted the correct GO term 60% to 70% of the time, with the remainder of the predictions being to insufficiently specific GO terms. The high precision was found to be due to the basing of electronic annotations on manually curated mapping files. Curators noted that it was more important for database curation to be accurate than to have complete coverage, and the figures above demonstrate that this is the tendency with electronic annotation.

Currently, we only receive manual annotations for a few plant species (a very large range of species are electronically annotated.) The current range of manual plant species annotation provides an excellent basis for comparison of plant gene products with the animal and bacterial gene products that have been annotated. However, the GO would be of greater utility if a larger range of plant species were also manually annotated. Annotation of the gene products from the full range of molecular biology model species is a specific goal of the GO Consortium, and we are very actively seeking to make links with groups that would be interested in making such annotations. Annotations could be provided either by established species-specific databases or by individual biologists (in cases where the community of interested researchers is too small to support a formal bioinformatics database structure). To contribute annotations, established databases should contact the GO Consortium directly via our mailing list, while individual biologists should contact the member database that is most appropriate to their model species. It is necessary for annotations to be handled by a member database resource since they will have the facilities to verify the initial annotations and to maintain them as the GO develops over time.



1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
TAIR	gene:1945249	PAD3		GO:0009700	TAIR:Publication:4743|PMID:8090752	IMP		P	PHYTOALEXIN DEFICIENT 3	AT3G26830	gene	taxon:3702	20020516	TAIR
TAIR	gene:1945257	PDS2		GO:0004161	TAIR:Publication:501708368	NAS		F	PHYTOENE DESATURATION 2	PDS2	gene	taxon:3702	20030912	TAIR
TAIR	gene:1944470	SEP3		GO:0005515	TAIR:Publication:501680061|PMID:11439126	IPI		F	SEPALLATA3	SEPALLATA3	gene	taxon:3702	20030312	TAIR
GR	Q01883	RA17		GO:0004867	GR_REF:1733|PMID:1376283	IEP		F			protein	taxon:4530	20041027	GR
GR	Q6I578	OSJNBb0014K18.3		GO:0003677	InterPro:IPR001584	IEA		F	Putative polyprotein		protein	taxon:39947	20041027	SPTR
GR	Q01883	RA17		GO:0004867	GR_REF:98	IEP		F			protein	taxon:4530	20041027	GR
SGD	S000002849	APT2	NOT	GO:0003999	SGD_REF:S000055423|PMID:9864350	IDA		F		YDR441C	gene	taxon:4932	20020902	SGD
SGD	S000005197	TEX1		GO:0006406	SGD_REF:S000069956|PMID:11979277	IC	GO:0000346	P		YNL253W	gene	taxon:4932	20030221	SGD

Figure 3. The gene_association file format. The gene_association file format is used for submission of annotations to the GO Consortium. An excerpt from a gene_association file format is shown, with eight individual annotations taking up one line each. The file is split half way along its horizontal length to allow it to fit in the page (break indicated by zigzag line). The content of the columns are explained in Table III and the evidence codes are explained in Table IV.

For further information and assistance with submitting annotations, contact the GO Consortium (see "Contacting GO," below).

GO AND PLANT SCIENCE

Some of the major areas currently of interest in plant biology are physiology, metabolism, inter- and intracellular transport, and signal transduction, as well as studies of whole plant growth and development, responses to changes in the environment, study of interactions between plants and their pests and pathogens, soil and rhizosphere biology, and understanding and exploitation of the diversity of plant form and function (Biotechnology and Biological Sciences Research Council Web site, http://www.bbsrc.ac.uk/science/areas/pms.html). Some of the processes being studied in plants are common with those being studied in nonplant species, while some are very widely different, and this is reflected within the GO.

The vast majority of terms within the GO that are used for plant annotation are shared with many other species. These include the transport terms, the metabolism terms such as those involved in respiration, and carbohydrate and amino acid processing. The same applies with the catalytic enzyme functions and the majority of the subcellular component terms.

In addition to this shared corpus of terms, there are a large number that have been added to accommodate the annotation of plant species, such as the terms developed by the PAMGO group discussed above. Many more plant-specific terms have also been added, including those covering flower and root development, pollination, and the C4 photosynthesis and crassulacean acid metabolism photosynthesis terms.

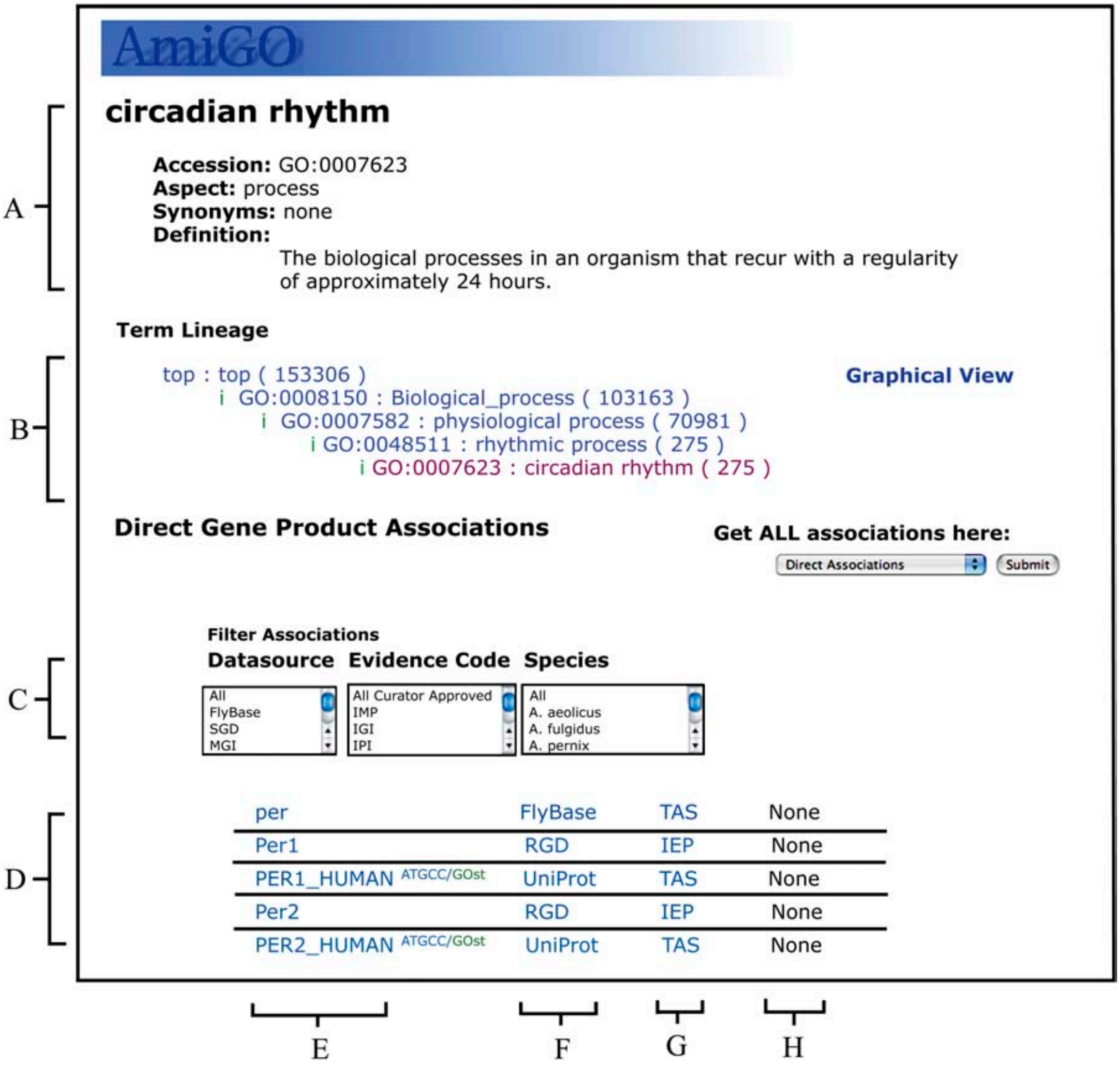
Some terms that were already added to the GO for the initial few species being annotated have had to be modified to accommodate annotation of plant gene products. For example, there were terms covering the generation of cell wall components in yeast (*Saccharomyces cerevisiae*) but these had to be modified to accommodate plant cell walls. Terms have also had to be modified to accommodate the differences between bacterial photosynthesis and plant photosynthesis, and between the systems associated with bacterial Rubisco and plant Rubisco. In some cases, this required use of sensu designation, as discussed earlier in the paper. The principal sensu designations of interest to plant biologists are Viridiplantae and Magnoliophyta. For the full list of sensu designations, see http://www.geneontology.org/GO.usage.shtml#taxon.

Although the GO includes a great many processes that do not occur in plants, it does not follow that annotations to these processes are irrelevant to plant biologists. For example, there is a range of terms for the development of structures that are not present within plants. The developmental processes required for the generation of these structures are often common with plant developmental processes. As a result, it is possible for plant scientists to use the information about gene products annotated to these terms to form hypotheses about the developmental processes in their model species. This principle applies with many of the other GO terms covering processes not occurring in plants.

The most extensive manually annotated species are baker's or budding yeast, mouse, and fruitfly since the database resources for these species were founding members of the Consortium. These species give excellent coverage of a range of processes and functions of interest to plant scientists. For example, ion transporters and metabolism terms are very well annotated and these annotations are of great use to plant biologists, especially those derived from yeast where these topics are extensively studied and where the processes will be closely related to their counterparts in plants. To see the levels of manual annotations currently achieved by each contributing database, go to http://www.geneontology.org/GO.current.annotations.shtml.

GO SOFTWARE

The GO Consortium makes available two tools to maximize the accessibility of the ontologies. The first is the Web-based GO browser AmiGO (http://www.godatabase.org/), shown in Figure 4. It displays the ontologies in an expandable tree view with relationship types shown. The database name associated with a given annotation is hyperlinked to the Web site of the database that submitted the annotation. The primary paper on which the annotation was based may be found on the submitting database's Web site. The AmiGO tool allows the kind of search that was ex-



circadian rhythm

Accession: GO:0007623
Aspect: process
Synonyms: none
Definition:
The biological processes in an organism that recur with a regularity of approximately 24 hours.

Term Lineage

top : top (153306)
i GO:0008150 : Biological_process (103163)
i GO:0007582 : physiological process (70981)
i GO:0048511 : rhythmic process (275)
i GO:0007623 : circadian rhythm (275)

Graphical View

Direct Gene Product Associations

Get ALL associations here:
Direct Associations
Submit

Filter Associations

Datasource	Evidence Code	Species
All FlyBase SGD MGI	All Curator Approved IMP IGI IPI	All A. aeolicus A. fulgidus A. pernix

per	FlyBase	TAS	None
Per1	RGD	IEP	None
PER1_HUMAN ATGCC/GOst	UniProt	TAS	None
Per2	RGD	IEP	None
PER2_HUMAN ATGCC/GOst	UniProt	TAS	None

Figure 4. AmiGO. AmiGO is a free online tool for searching the ontologies and manual annotations (http://www.godatabase.org/). A term is displayed, with its name, definition, unique identifier and synonyms (A). Following this, the position in the ontology is displayed (B) with the relationships marked as i for is_a and p for part_of. The gene product association search may be refined by database, evidence code or species (C), and the search results are displayed at the bottom of the page (D). Each row begins with the gene product symbol for a gene product that has been associated with the GO term displayed above (E). Following this is a link to the entry for the gene product in the submitting database resource (F). Primary papers providing evidence for the association of the gene product to the particular GO term may be found on the submitting database resource Web site. The evidence codes (G) and the full gene or gene product names follow (H).

plained at the beginning of this paper. As a straightforward example, a scientist could search for information on gene products involved in circadian rhythms in plants. In the search box, they would type circadian rhythm and they would choose to search all GO terms. The search result would produce all the related terms and these would be hyperlinked to a listing showing the expanded ontology structure. On the page with the expanded structure, there would also show the listing of all gene products that have been annotated to the given process. In the more complex situation described at the beginning of the paper, a scientist might search for the ambiguous concept "gametogenesis." A keyword literature search on this term would pull out all papers covering plant gametophyte development and animal gamete production. However, a search in AmiGO on this term would draw out several process names, including male gametophyte development and female gametophyte development. The definitions of all the relevant processes would also be displayed so that the user could decide which is the process of interest. The term names would be hyperlinked to the expanded ontology structure showing the process of interest. Individual gene product names would provide further links to publications showing specific evidence of the involvement of the annotated gene products in the process described. GO browsers like AmiGO allow scientists to find information on gene products involved in given processes, across a range of species. They remove the difficulties in searching that could be caused by ambiguous technical language and therefore open up the literature of unfamiliar fields for full investigation.

The Consortium also produces a custom ontology editor called OBO-Edit. The application can be used to display the ontologies and to browse through them. It displays the annotated gene products via a plug-in and allows creation and editing of new ontologies and of slims (described below) from existing ontologies.

In addition to the Consortium tools, a range of other third-party Web-based and standalone tools are available. See the GO Tools page for further details (http://www.geneontology.org/GO.tools.shtml).

GO SLIMS

In addition to making the full ontologies and annotation sets available, the Consortium provides a range of custom datasets. The Consortium has prepared GO slims, slimmed down versions of the ontologies that allow you to annotate genomes or sets of gene products to gain a high-level overview of gene functions. GO slims are versions of the ontologies in which the more specific terms (and therefore their annotations) have been collapsed up into the more general parent terms; for example, "style development" can be collapsed into "flower development." The GO Consortium maintains both a generic and a plant-specific GO slim. Slims can be particularly useful for a group wishing to use just a subsection of the GO for analysis in a particular field or of a particular subset of a genome. Using GO slims, a scientist can, for example, work out the proportion of a genome that is involved in signal transduction, biosynthesis, or reproduction. (The accuracy of this method is dependent on the status of the annotation of a given species.) Groups can create their own GO slims and then the annotations can be fitted to the slim using a perl script provided by the Consortium (map2slim.pl). More information on GO slims can be found at http://www.geneontology.org/GO.slims.shtml.

CROSS-REFERENCES

GO is not the only attempt to build structured controlled vocabularies for genome annotation. Nor is it the only such series of catalogs in current use. For example, newcomers to GO may already be familiar with the longer established protein product nomenclature, the Enzyme Commission Enzyme Nomenclature (http://www.chem.qmul.ac.uk/iubmb/enzyme/). The GO draws heavily on such preexisting systems, and we provide translation tables (mappings) that list the cross-references between these other catalogs and GO. A mapping file is created by finding analogous concepts in two catalogs and listing in a file the concepts that correspond. For example, if an enzyme function is represented in GO and that enzyme is also listed in the Enzyme Commission database, then we would make a cross-reference as a term "general dbxref" in the GO term entry. We would then list the corresponding concepts in a text file and call this the ec2go mapping file. Such files are useful for users who wish to transfer gene product annotations from other resources to the GO. We do not include every single E.C. numbered enzyme function in the GO but instead we add the function terms as they are required by our annotators. If we are asked to add an enzyme function term that does not have an E.C. number, then we will add that term and then an E.C. number can be added later when it becomes available. The GO does not seek to supersede the E.C. system, but compliments it, since the two systems are developed for different purposes and since the GO is fully cross-referenced to the enzyme nomenclature database.

Note that while our mappings are of high quality, they are neither complete nor exact. More information on the syntax of these mappings can be found in the GO File Format Guide, which is available online. In addition to these mappings, GO is included as a source vocabulary in the National Library of Medicine's (NLM's) Unified Medical Language System (UMLS), which includes Medical Subject Headings terms.

OBTAINING GO DATA

All GO tools and resources are free of charge and open source and can be downloaded from the GO Web site. Users can download the ontology files in four different formats: OBO flat files (updated daily), GO flat files (updated daily, an older format, still supported, but not recommended for use), XML (updated daily), and MySQL (updated weekly). For more information on the syntax of these formats, see the GO File Format Guide http://www.geneontology.org/GO.format.shtml.

OBO

GO allows annotation of genes and their products with a limited set of attributes (process, function, component). Annotation of other attributes requires other ontologies, such as the range available on the Open Biomedical Ontologies (OBO) Web site (http://obo.sourceforge.net/). OBO is a GO Consortium Web-based repository for ontologies developed outside of the Consortium. There are a small number of criteria for inclusion of an ontology in OBO. The ontologies must be open source. They must be able to be used by all, without any constraint other than that their origin must be acknowledged, and that they cannot be altered and redistributed under the same name. All OBO ontologies must be in, or must be able to be instantiated in, a common shared syntax. The GO syntax, extensions of this syntax, and Web Ontology Language are all suitable. Ontologies in OBO must be orthogonal to one another and must each have a unique identifier space. Finally, all terms within the recently added ontologies must include textual definitions of their terms.

The GO Consortium supports the development of the ontologies in OBO and makes its tools for editing and curating freely available. Ontologies of particular interest to plant scientists will include the plant growth and developmental stage ontologies and the plant structure ontology produced by the Plant Ontology Consortium (POC, http://www.plantontology.org/) for gene expression and phenotype annotation. The plant growth stage ontology, which is currently under development, will describe stages in the growth and development of plants, including the development of individual organs. Terms will complement but will not overlap with developmental GO process terms such as flower development and ovule development. The OBO ontologies can be used in combination with the GO ontologies for full annotation of gene products, and those that can be used in combination to produce cross-product terms are marked as such on the OBO Web site (Hill et al., 2002). If you are using other ontologies from the OBO site, it would be useful to bear in mind that these are developed by groups outside of the GO Consortium, and that the rules used in their development, and the caveats applying in their use may differ from those mentioned here with respect to the GO Consortium ontologies. The rules stated above are the only criteria for inclusion of ontologies within OBO and the ontologies are not checked in any way beyond this by the GO Consortium. Users should request further information from the developers, whose contact details are listed on the Web site, and decide for themselves whether the ontologies will suit their needs before using them. Some useful information to have before beginning use of an ontology includes how long it has been in development, and what the arrangements for feedback and improvement of the ontology are. It is also useful to know how long the ontology will continue to be supported after its initial production, and whether the ontology has previously been used in combination with other ontologies, if you wish to do this.

CONTACT GO

To support the continued development of GO, the Consortium continually seeks contact with biological researchers and bioinformatics groups so that consensus across the biological community may be achieved. We welcome questions, requests, and ideas from our users and are especially keen that biologists should participate in GO's development. Requests for new terms may be submitted via the mailing list (go@geneontology.org) or the sourceforge tracker (http://geneontology.sourceforge.net/), and the discussions leading to changes to the GO may be viewed online in the sourceforge tracker and in the e-mail archives (http://www.geneontology.org/GO.contents.archives.mail.shtml). The GO Consortium also runs GO Users Meetings, where users can meet Consortium members and other GO users in person. These meetings are open to anyone interested in the GO project, and provide opportunities for GO users and developers of GO-related analysis and visualization tools to share their work with each other and with GO Consortium members.

ACKNOWLEDGMENTS

We thank Judith Blake, Michael Ashburner, Midori Harris, Amelia Ireland, Emily Dimmer, Elisabeth Truernit, Fiona McCarthy, Daniel Barrell, Tanya Berardini, and Anuradha Pujar for constructive comments on the manuscript.

Received December 17, 2004; revised April 1, 2005; accepted April 1, 2005; published July 11, 2005.

LITERATURE CITED

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet **25:** 25–29

Berardini TZ, Mundodi S, Reiser L, Huala E, Garcia-Hernandez M, Zhang P, Mueller LA, Yoon J, Doyle A, Lander G, et al (2004) Functional annotation of the Arabidopsis genome using controlled vocabularies. Plant Physiol **135:** 745–755

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell **16:** 1679–1691

Camon E, Magrane M, Barrell D, Binns D, Fleischmann W, Kersey P, Mulder N, Oinn T, Maslen J, Cox A, et al (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. Genome Res **13:** 662–672

Camon EB, Barrell DG, Dimmer EC, Lee V, Magrane M, Maslen J, Binns D, Apweiler R (2005) An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. BMC Bioinformatics (Suppl 1) **6:** S17

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, et al (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science **296:** 92–100

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, et al (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res **32:** D258–D261

Hill DP, Blake JA, Richardson JE, Ringwald M (2002) Extension and integration of the Gene Ontology (GO): combining GO vocabularies with external vocabularies. Genome Res **12:** 1982–1991

Li R, Rimmer R, Buchwaldt L, Sharpe AG, Seguin-Swartz G, Coutu C, Hegedus DD (2004) Interaction of Sclerotinia sclerotiorum with a resistant Brassica napus cultivar: expressed sequence tag analysis identifies genes associated with fungal pathogenesis. Fungal Genet Biol **41:** 735–753

Ma X, Dong Y, Matzuk MM, Kumar TR (2004) Targeted disruption of luteinizing hormone beta-subunit leads to hypogonadism, defects in gonadal steroidogenesis, and infertility. Proc Natl Acad Sci USA **101:** 17294–17299

Malek RL, Sajadi H, Abraham J, Grundy MA, Gerhard GS (2004) The effects of temperature reduction on gene expression and oxidative stress in skeletal muscle from adult zebra fish. Comp Biochem Physiol C Toxicol Pharmacol **138:** 363–373

Mi H, Vandergriff J, Campbell M, Narechania A, Majoros W, Lewis S, Thomas PD, Ashburner M (2003) Assessment of genome-wide protein function classification for Drosophila melanogaster. Genome Res **13:** 2118–2128

Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, et al (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature **420:** 563–573

Oliver MJ, Dowd SE, Zaragoza J, Mauget SA, Payton PR (2004) The rehydration transcriptome of the desiccation-tolerant bryophyte Tortula ruralis: transcript classification and analysis. BMC Genomics **5:** 89

Pouliot Y, Gao J, Su QJ, Liu GG, Ling XB (2001) DIAN: a novel algorithm for genome ontological classification. Genome Res **11:** 1766–1779

Robertson WR, Clark K, Young JC, Sussman MR (2004) An Arabidopsis thaliana plasma membrane proton pump is essential for pollen development. Genetics **168:** 1677–1687

Silva NF, Goring DR (2002) The proline-rich, extensin-like receptor kinase-1 (PERK1) gene is rapidly induced by wounding. Plant Mol Biol **50:** 667–685

Xie H, Wasserman A, Levine Z, Novik A, Grebinskiy V, Shoshan A, Mintz L (2002) Large-scale protein annotation through Gene ontology. Genome Res **12:** 785–794

Zdobnov EM, Apweiler R (2001) InterProScan: an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17:** 847–848

Zhou Y, Zhou C, Ye L, Dong J, Xu H, Cai L, Zhang L, Wei L (2003) Database and analyses of known alternatively spliced genes in plants. Genomics **82:** 584–595