The Plant Cell, Vol. 20: 1426–1436, June 2008, www.plantcell.org © 2008 American Society of Plant Biologists

CURRENT PERSPECTIVE ESSAY
Special Series on Large-Scale Biology

Arabidopsis Reactome: A Foundation Knowledgebase for Plant Systems Biology[W]

Nicolas Tsesmetzis,[a] Matthew Couchman,[a] Janet Higgins,[a] Alison Smith,[b] John H. Doonan,[c] Georg J. Seifert,[c] Esther E. Schmidt,[d] Imre Vastrik,[d] Ewan Birney,[d] Guanming Wu,[e] Peter D'Eustachio,[e,f] Lincoln D. Stein,[e] Richard J. Morris,[a] Michael W. Bevan,[c,1] and Sean V. Walsh[a,1]

[a] Department of Computational and Systems Biology, John Innes Centre, Norwich NR4 7UH, United Kingdom
[b] Department of Metabolic Biology, John Innes Centre, Norwich NR4 7UH, United Kingdom
[c] Department of Cell and Developmental Biology, John Innes Centre, Norwich NR4 7UH, United Kingdom
[d] European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD, United Kingdom
[e] Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724
[f] New York University, School of Medicine, New York, New York 10016

New ways of capturing and representing biological knowledge are needed to enable individual researchers to remain abreast of relevant discoveries and to permit computational approaches for interpreting the large volumes of diverse data generated by modern biological research. Here, we describe a promising approach that expands the term ''reaction'' to represent biological processes. We show how users can represent a wide variety of biological processes in plants in terms of the concept of a reaction and assemble the information obtained from the model plant *Arabidopsis thaliana* into an online knowledgebase called Arabidopsis Reactome. Its curated and imported pathways currently cover ~8% of the *Arabidopsis* proteome. Arabidopsis Reactome events have also been electronically projected onto five other predicted plant proteomes. Such a system allows the visualization and interpretation of high-throughput data, hypothesis formulation in systems biology, and is a useful learning resource. The Arabidopsis Reactome project (www.arabidopsisreactome.org) is open access, open source, and open to contributions.

OVERVIEW

Currently, the genome sequences of six higher plants and a moss species have been assembled, annotated, and published (Arabidopsis Genome Initiative, 2000; International Rice Genome Sequencing Project, 2005; Tuskan et al., 2006; Jaillon et al., 2007; Velasco et al., 2007; Ming et al., 2008; Rensing et al., 2008). The availability of large populations of sequence-tagged insertion mutations for nearly all *Arabidopsis* genes (Alonso and Ecker, 2006), surveys of polymorphisms in many *Arabidopsis* ecotypes (Clark et al., 2007), and the free availability of microarray data and data-mining tools (Craigon et al., 2004; Zimmermann et al., 2004) have all greatly accelerated the scale and scope of plant research (Somerville and Koornneef, 2002), reflected by more than 240 publications per month citing *Arabidopsis*. Given the acceleration in the number of plant genome sequencing projects, it is increasingly difficult for individual researchers to stay abreast of relevant literature and to make connections between different sets of information. This difficulty is compounded by the general inaccessibility of information contained in the literature to computer-based analysis, which severely limits its value as a source of biological knowledge (Jensen et al., 2006). Therefore, developing computational methods for capturing and representing biological knowledge is a high priority, particularly for model organisms that are the focus of most experimental work.

Several bioinformatics resources and software packages have been developed to manage and exploit the wealth of data generated by plant genome projects, functional genomics resources, and high-throughput transcriptomics experiments. The predicted proteins in the *Arabidopsis* genome have been systematically described by The Arabidopsis Information Resource (TAIR; http://www.arabidopsis.org/portals/genAnnotation/) using Gene Ontology (GO)–controlled descriptions of gene functions according to the biological process, molecular function, and cellular component of individual genes (Ashburner et al., 2000). This has enabled much more rapid, accurate, and consistent assignment of predicted functions to genes and permits the development of more accurate relationships between genes in different organisms. Several databases and software applications relate gene entities to each other in networks in *Arabidopsis*. The AraCyc database (Mueller et al., 2003) displays computationally predicted *Arabidopsis* metabolic pathways that are largely manually curated. MAPMAN uses a hierarchical ontology different from GO terms that can be used for visualizing large data sets onto metabolic pathways and other biological processes (Thimm et al., 2004). The VirtualPlant (Gutierrez et al., 2007) and ONDEX (Kohler et al., 2006) systems have created graph-based integrations of knowledge and gene functional inferences that may be queried, filtered, and appended using

[1] Address correspondence to michael.bevan@bbsrc.ac.uk or sean.walsh@bbsrc.ac.uk.
[W] Online version contains Web-only data.
www.plantcell.org/cgi/doi/10.1105/tpc.108.057976

tools like Cytoscape (Suderman and Hallett, 2007) to generate new functional insights. GENEVESTIGATOR (Zimmermann et al., 2004, 2005) provides web-based analytical services that relate gene expression data to a wide variety of gene-related entities, such as GO terms, mutant phenotypes, pathways, and developmental processes.

Reactome is an extensively curated pathway knowledgebase that focuses on human processes (Joshi-Tope et al., 2003, 2005; de Bono et al., 2007; Vastrik et al., 2007). A key feature of Reactome is its elegant data model that extends the notion of a biochemical reaction, where substrates go in, products come out, and a catalyst is frequently required to lower the free energy of the transformation. This concept also can be used to represent the binding of a ligand to a membrane receptor, the formation of a complex, the binding of a transcription factor in a promoter region, or the translocation of a molecule between subcellular compartments. In this way, the data model expresses molecular processes in the same way that scientists understand them and allows connected reactions to represent biological processes (e.g., transcription and cell cycle) in terms of their underlying molecular transformations, associations, and translocations. Based on this extended definition of reaction, reactants and products can be proteins, lipids, nucleotides, small molecules, or complexes of these. The data model further distinguishes among the topologically or functionally different forms of each molecule. For instance, this allows the distinction between chloroplastic maltose and cytosolic maltose or between the various post-translational modifications of a protein.

Here, we describe Arabidopsis Reactome, a knowledgebase of biological processes from the model plant *Arabidopsis.* Release 2 (www.arabidopsisreactome.org) comprises seven curated and 311 imported superpathways that together represent 8% of the *Arabidopsis* proteome. We show, using examples based on the mitotic cell cycle, that the knowledgebase has wide applicability for exchanging structured data with other databases, for comparative network analysis, data integration, and for visualization and protein interaction analysis. The straightforward authoring tool and realistic detailed descriptions of biological processes inherent in Reactome's data model provide an excellent foundation for representing, exchanging, and integrating biological information, suggesting that it will find wide application in the *Arabidopsis* community as a gold standard for pathway knowledge and key foundation for systems biology research.

PATHWAY CURATION

The information in Arabidopsis Reactome was generated from curated pathways that have been manually entered and reviewed by experts and imported pathways from several third-party pathway databases. Knowledge acquisition for the curated pathways followed the process established for the human Reactome system. Essentially, pathways were authored, curated, and peer reviewed by expert biologists (PhD level and above) and bioinformaticians. The curatorial process used a set of applications, namely, Reactome Author Tool and Reactome Curator Tool, developed specifically for the purpose of collecting and validating pathway models (Joshi-Tope et al., 2005). Every protein, gene, or small molecule in Arabidopsis Reactome has a reference identifier that points into a public reference database. In the case of protein sequences, the primary source of identifiers is UniProt. Entities such as chemical compounds are referenced by the ChEBI database. For the imported KEGG and AraCyc enzymes and chemical compounds, referencing was performed in an automated manner that is also applicable to newer additions. In addition, entries were automatically cross-referenced to external databases, such as UniProt (Schneider et al., 2005), TAIR (Rhee et al., 2003), Munich Information Center for Protein Sequences (MIPS) (Schoof et al., 2004), National Center for Biotechnology Information (NCBI) Entrez Gene (Maglott et al., 2005), KEGG COMPOUND (Kanehisa and Goto, 2000), and ChEBI (Degtyarenko et al., 2007).

CURATED PATHWAYS

The long-term goal in Arabidopsis Reactome is to establish a detailed set of curated pathways representing all major biological processes in *Arabidopsis*. Initially, metabolic pathways and the mitotic cell cycle were selected as contrasting processes that would challenge the plasticity of the data model and provide a foundation for data integration and modeling. An essential piece of information for a reaction to be incorporated into the curated area of Arabidopsis Reactome was the existence of experimental evidence, usually a reference to a published article. In addition to reactions and literature references, the data model contains fields for species, GO molecular function, subcellular location, and other relevant information that were filled out during the curatorial process. In some instances, *Arabidopsis* reactions imported from KEGG and AraCyc databases were used as structured reference material to start the curatorial process.

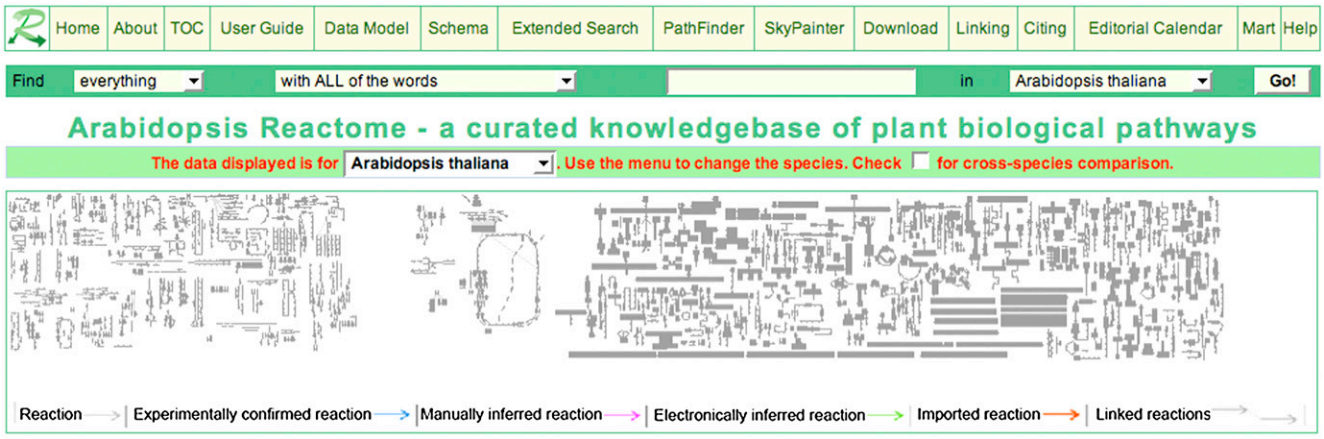
When there was insufficient experimental evidence of a particular reaction in *Arabidopsis* but its existence could be inferred (for example, gene functions inferred from sequence similarity), then the equivalent reaction could be manually inferred from a different organism from which there was sufficient experimental evidence. For example, we deduced the binding of the five-subunit Replication factor C (RFC) onto DNA that results in the displacement of the polymerase α (POLA) on the basis of the experimentally confirmed activity in human and by the identification in *Arabidopsis* of all five RFC subunits containing the characteristic sequence motifs of other eukaryotic RFCs, as described by Shultz and Furukawa (Furukawa et al., 2003; Shultz et al., 2007). In the Arabidopsis Reactome user interface, manually inferred reactions are flagged in magenta to distinguish them clearly from the curated ones, which appear in blue. Once new experimental evidence becomes available in *Arabidopsis* for any of the inferred reactions, these reactions can be replaced by the experimentally determined examples during a scheduled pathway review and then

CURRENT PERSPECTIVE ESSAY

will appear as blue arrows in the next public release. Inferred reactions are used sparingly so as not to jeopardize the integrity of a pathway. All curated and manually inferred reactions occupy the central part of the reaction map on the Arabidopsis Reactome home page (Figure 1).

IMPORTED PATHWAYS

We imported *Arabidopsis* metabolic pathways from KEGG (release 38.0) and AraCyc (release 2.5) databases into Arabidopsis Reactome as text files from their ftp servers. The files were parsed, and the data were stored in a MySQL relational database using custom database schemata developed to represent each source. Using the Perl XML::Generator module (http://www.cpan.org/), these data were used to construct documents in the Reactome Author Tool native, XML-based, *GKB* format. The documents were then opened with the Author Tool, and reactions were joined manually to form pathways according to the pathway diagrams found at their source websites. These files were then imported into the Reactome Curation Tool that was used to deposit the data in the central Arabidopsis Reactome database. Once in the database, the arrows representing the reactions were manually laid out in the reaction map using the Reactome pathway visualization tool. Pathways involved in related or similar processes were laid out in close proximity to each other within either the KEGG or



Cell Cycle, Mitotic	**Cell Cycle Checkpoints**	**DNA Replication**	**Endoreduplication and Alternative Cell Cycles**
Nucleotide Sugar Metabolism	**Starch degradation in leaves**	**Sucrose Biosynthesis**	Alanine and aspartate metabolism (KEGG)
Alkaloid biosynthesis I (KEGG)	Alkaloid biosynthesis II (KEGG)	Aminophosphonate metabolism (KEGG)	Aminosugars metabolism (KEGG)
Arachidonic acid metabolism (KEGG)	Arginine and proline metabolism (KEGG)	Ascorbate and aldarate metabolism (KEGG)	Benzoate degradation via CoA ligation (KEGG)
Benzoate degradation via hydroxylation (KEGG)	beta-Alanine metabolism (KEGG)	Biosynthesis of steroids (KEGG)	Biotin metabolism (KEGG)
Butanoate metabolism (KEGG)	C21-Steroid hormone metabolism (KEGG)	Carbon fixation (KEGG)	Citrate cycle (TCA cycle) (KEGG)
Cyanoamino acid metabolism (KEGG)	Cysteine metabolism (KEGG)	Diterpenoid biosynthesis (KEGG)	Fatty acid biosynthesis (KEGG)
Fatty acid metabolism (KEGG)	Flavonoid biosynthesis (KEGG)	Folate biosynthesis (KEGG)	Fructose and mannose metabolism (KEGG)
Galactose metabolism (KEGG)	gamma-Hexachlorocyclohexane degradation (KEGG)	Ganglioside biosynthesis (KEGG)	Glutamate metabolism (KEGG)
Glutathione metabolism (KEGG)	Glycerolipid metabolism (KEGG)	Glycerophospholipid metabolism (KEGG)	Glycine, serine and threonine metabolism (KEGG)
Glycolysis / Gluconeogenesis (KEGG)	Glycosphingolipid metabolism (KEGG)	Glyoxylate and dicarboxylate metabolism (KEGG)	Histidine metabolism (KEGG)
Indole and ipecac alkaloid biosynthesis (KEGG)	Inositol metabolism (KEGG)	Inositol phosphate metabolism (KEGG)	Limonene and pinene degradation (KEGG)
Lipopolysaccharide biosynthesis (KEGG)	Lysine biosynthesis (KEGG)	Lysine degradation (KEGG)	Metabolism of xenobiotics by cytochrome P450 (KEGG)

Figure 1. Overview of the Arabidopsis Reactome Home Page.

The panel with the arrows is the reaction map. The arrows represent reactions that have been manually created based on evidence from the literature and reactions that have been imported from external databases (AraCyc and KEGG). Below the reaction map is the table of modules that contains a list of all the superpathways that are present in Arabidopsis Reactome.

AraCyc areas of the reaction map. Importation software has been written to allow Arabidopsis Reactome to be updated from KEGG and AraCyc sources as new versions become available.

Our analyses identified important differences between pathways represented in the KEGG and AraCyc databases that allowed us to improve data quality. Originally, these pathways were computationally predicted for the sequenced *Arabidopsis* genome using inference methods from GENES and MetaCyc databases, respectively (Kanehisa et al., 2002; Mueller et al., 2003). Since then, many AraCyc pathways have been manually curated. One of the consequences of the absence of manual curation of the *Arabidopsis* data in the KEGG database is that key enzymes, such as sucrose-6-phosphate phosphatase (SPP), do not appear on KEGG reaction diagrams. Similarly, although AraCyc describes the four SPP isoforms, it relies on computational inference and not literature references. Finally, SPPs are also absent from the *Arabidopsis* inferred part of the human Reactome since humans do not synthesize sucrose.

SUBA (Heazlewood et al., 2007) and UniProt subcellular localization information was used to assign the subcellular location of the imported reactions. For example, the enzyme peroxisomal 2,4-dienoyl-CoA reductase was used to assign its catalyzing reaction ''monovinyl protochlorophyllide a + NADPH <=> $NADP^+$ + chlorophyllide a [AT3G12800]'' and its components to the peroxisome. This substantially enriches the knowledge associated with metabolic pathway data.

ELECTRONIC INFERENCE OF *ARABIDOPSIS* PATHWAYS ONTO OTHER PLANTS

The predicted proteomes from the published genomes of rice (*Oryza sativa*; International Rice Genome Sequencing Project, 2005), poplar (*Populas trichocarpa*; Tuskan et al., 2006), the moss *Physcomitrella patens* (Rensing et al., 2008), and the two grape varieties (*Vitis vinifera* and *V. vinifera* var *Pinot Noir*; Jaillon et al., 2007; Velasco et al., 2007) were downloaded from their websites. Using the NCBI BLASTP algorithm, we matched the *Arabidopsis* proteome downloaded from the TAIR ftp site to the predicted proteomes of these species and then fed the results to the OrthoMCL algorithm (Li et al., 2003) to identify and cluster the orthologous proteins into groups. OrthoMCL results were then appropriately formatted so they could be used by the scripts included in the Reactome system to produce the equivalent organism-specific reactions and pathways.

A total of 8269 reactions and 2196 proteins from *Arabidopsis* were projected onto the five plant species (Table 1) to identify evolutionary conserved pathways. These projected pathways can be used for overlaying ''-omics'' data, cross-species comparisons of biological processes, rapid proteome annotation of a newly sequenced plant genome, and to provide a template for the establishment of other species-specific Reactome knowledgebases.

Comparison of the *Arabidopsis* cell cycle with the electronically inferred cell cycle in rice, poplar, *P. patens*, and the two grape varieties showed that almost 60% of the reactions were conserved between *Arabidopsis* and poplar, compared with 45% in rice and 33% in moss (Table 2). In the case of grape, the two genomes showed different levels of conserved cell cycle reactions (51 and 39%). This may be attributable to different gene finding methods or problems in genome assembly where the genome is highly heterozygous. However, both S and M phases were more conserved in all species compared with the G1, G0, and G2 phases and their transitions. These data can be seen on the Arabidopsis Reactome website when the cross-species comparison box is ticked. By revealing apparent similarities and differences across species, Arabidopsis Reactome identifies gaps in pathways that can be used to either confirm a different molecular basis for equivalent phenomena or to improve existing gene models in light of comparative evidence.

THE USER INTERFACE

The Arabidopsis Reactome home page is divided into two main panels: the reaction map and the table of contents (Figure 1). The reaction map displays all the reactions contained in Arabidopsis Reactome in the form of arrows. Arrows joined together represent biological pathways, and pathways are clustered according to related biological processes. The left side of the reaction map is occupied by *Arabidopsis* reactions found in KEGG (Kanehisa et al., 2002), and the right side is taken by those reactions found in AraCyc (Mueller et al., 2003), leaving the center for those reactions that have been manually curated. The table of contents is seen under the reaction map. It lists all the primary pathways (superpathways) present in Arabidopsis Reactome, starting with the curated pathways at the top of the table and followed by pathways from KEGG and AraCyc. As curated, peer-reviewed pathways are added to the central part of the reaction map, equivalent pathways from KEGG and AraCyc are removed, increasing the number of reactions in the central panel and decreasing those from the side panels.

Entries in Arabidopsis Reactome can be searched using simple or advanced search facilities located at the top of the home page. By selecting a reaction, the web interface of Arabidopsis Reactome returns an increasing level of detailed information. This includes a description, the reaction components (compounds, enzymes, etc.), GO annotation (subcellular location and molecular function), preceding and following events, organism name, equivalent events in other organisms, reference in the literature, and, where applicable, links to external databases, such as UniProt, TAIR, MIPS, NCBI Entrez Gene, KEGG COMPOUND, and ChEBI (Figure 2).

DATA INTEGRATION AND ANALYSIS

A key feature of a curated knowledgebase is its utility for integrating diverse data sets into a comprehensive description of biological processes. SkyPainter is a useful feature of the Reactome system that allows researchers to visualize and analyze their own data sets in relation to the reaction maps. It

CURRENT PERSPECTIVE ESSAY

Table 1. Statistics from the Orthologous Transfer of *Arabidopsis* Reactions onto Five Other Sequenced Plant Genomes

Statistics of Arabidopsis Reactome (Release 2)					
Species	Proteins	Complexes	Reactions	Pathways	Conserved Reactions (%)
Arabidopsis	2,196	86	8,269	1,108	100
P. trichocarpa	3,112	53	6,426	1,079	78
O. sativa	2,150	47	4,808	1,055	58
V. vinifera	2,636	51	6,476	1,070	78
V. vinifera var *Pinot Noir*	6,196	33	5,917	1,046	72
P. patens	1,414	44	3,691	1,036	45
Total	17,704	314	35,587	6,394	–

Reactions and pathway figures include duplicated events found in more than one superpathway.

can be found on the top menu bar of the Arabidopsis Reactome home page. Researchers can upload a list of genes or other identifiers to color the reaction map in a number of ways. The SkyPainter module recognizes a range of gene identifiers, such as Arabidopsis Genome Initiative code, Affymetrix probe set, and UniProt ID. Plants with electronically inferred pathways can also be searched by selecting the plant species on the SkyPainter page and using the appropriate gene or sequence identifier. We demonstrate the use of SkyPainter using two sets of published experimental data: for the overrepresentation analysis of glucose upregulated genes (Li et al., 2006) and for the visualization of the progression of gene expression during the cell cycle (Menges et al., 2003).

In the first example, SkyPainter was used for the overrepresentation analysis of genes maximally induced at 6 h in response to glucose treatment (Figure 3). For the overrepresentation analysis, SkyPainter uses the hypergeometric test (Feller, 1968) to color pathways according to the statistical likelihood that they would contain the listed genes by chance. Pathways containing hits (i.e., gene IDs) at higher probability than expected by chance are listed in the output and ordered based on their probability values. The pathway list can be expanded to show the individual events and matching genes within the pathways (Figure 3). Six hours after glucose treatment, the mitotic cell cycle pathway is the most highly overrepresented. Pathways such as DNA replication, choline biosynthesis, and glycolysis are also significantly overrepresented in response to glucose treatment. In addition to the statistical analysis, SkyPainter output shows the complete reaction map, with the reaction arrows colored according the number of genes in the submitted list that participate in the reaction. The color of arrows within the cell cycle pathway showed that reactions at the G1/S transition were most active 6 h after glucose treatment (Figure 3; see Supplemental Figure 1 online).

In the second example, SkyPainter was used to represent dynamic changes in expression of cell cycle genes throughout the cell cycle. Using publicly available microarray data available from NASCARRAYS (Craigon et al., 2004), we selected experiment NASCARRAYS-360 on genome-wide cell cycle studies. The experiment was designed to follow the expression of genes throughout the cell cycle in a synchronized *Arabidopsis* cell suspension. The cells were synchronized using aphidocolin, which upon removal allowed the resumption of the S phase and progression through the cell cycle (Menges et al., 2003). Samples were taken at 10 time points between 0 and 19 h after the removal of aphidocolin and analyzed using the Affymetrix ATH1 array (Menges et al., 2003). The expression data (normalized using Affymetrix MAS 5.0 scaling protocol) was downloaded, and the expression values of the 100 cell cycle genes that correspond to the genes curated within the cell cycle module of the Arabidopsis Reactome were extracted. These expression values were then used with Skypainter to create a movie that follows the quantitative expression levels of mitotic genes through the various phases of the cell cycle and reaching its peak at the 10-h time point (Figure 4; see Supplemental Movie 1 online). This provides a useful quantitative and visual tool for following gene expression patterns and supporting data interpretation.

The large increase in protein–protein interaction data (Cui et al., 2007; Geisler-Lee et al., 2007; Van Leene et al., 2007) provides an exceptionally rich source for extending known pathways and complexes. We integrated protein interaction data with the cell cycle pathway in Arabidopsis Reactome to identify new protein clusters associated with the cell cycle (see Supplemental Methods Online). The cell cycle data from

Table 2. Projection of the *Arabidopsis* Mitotic Cell Cycle Reactions onto Five Sequenced Plant Genomes

Species	Present	Absent	Conserved (%)
Arabidopsis	152	0	100
P. trichocarpa	86	66	57
V. vinifera	78	74	51
O. sativa	68	84	45
V. vinifera var *Pinot Noir*	60	92	39
P. patens	50	102	33

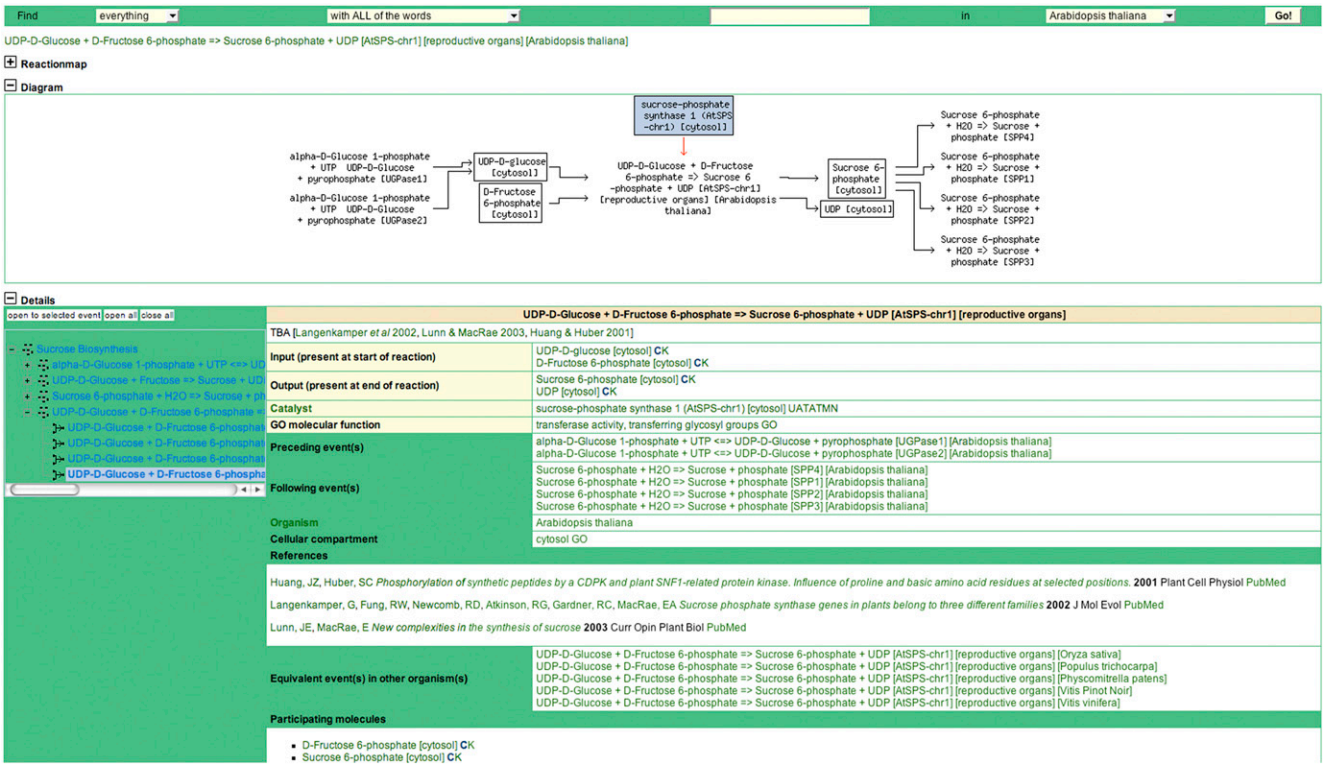


Figure 2. View of a Selected Reaction in Arabidopsis Reactome.

The image shows a diagrammatic representation of the reaction components and a list of information about the reaction, such as the inputs, outputs, catalyst, GO molecular function, preceding and following events, compartment, literature references, and equivalent events in other organisms. A hierarchical structure of the involved pathway can also be viewed by expanding the ''event hierarchy'' frame (shown on the left).

Arabidopsis Reactome was imported into Cytoscape in SBML format, and its proteins were joined with proteins from the experimental cell cycle interactome determined by protein mass spectrometry (Van Leene et al., 2007) to form a cell cycle network (see Supplemental Figures 2 and 3 online). We then enriched this network by importing the comprehensive set of predicted protein interactions from AtPID, which contains 11,708 proteins and 24,419 interactions (Cui et al., 2007) into Cytoscape and merged this data with the cell cycle network (Figure 5). Proteins common to the experimental and predicted interactomes were identified, and the first neighbors (proteins predicted to interact directly) of these nodes were identified in AtPID and added to extend the network (see Supplemental Methods online). The molecular complex detection (MCODE) plug-in (Bader and Hogue, 2003) was then used to detect potential clusters (see Supplemental Methods online). MCODE detects densely connected regions of protein interaction networks that may represent complexes (or protein families). Overall, 1201 new interactions related to the cell cycle have been identified. The proteins in seven of these clusters (numbered 1 to 7 in Figure 5) are shown in Supplemental Table 1 online, with the proteins curated in Arabidopsis Reactome and present in the cell cycle interactome identified. The predicted functions of proteins in these clusters are closely aligned with cell cycle regulation. For example, Cluster 1 is built around the Arabidopsis Reactome entities At MPK4 and At MPK6, two cytoplasmic mitogen-activated protein (MAP) kinases that phosphorylate MAP65-1 and inhibit microtubule bundling during spindle assembly in mitosis (Smertenko et al., 2006). Cluster 1 contains 12 other MAP kinases, including At MPK7 and mitogen-activated kinases that may link environmental responses to cell cycle control. Cluster 2 is based on the Arabidopsis Reactome entity NOC3, a nuclear protein that associates with ORC:origin to enable CDC6 interaction during initiation complex formation in the M/G1 transition. Cluster 2 contains a NOC2 homolog, several BRIX domain proteins and pescadillo-like protein predicted to be involved in rRNA processing and cell proliferation control, several GTP/RNA binding proteins, a crooked neck spliceosome protein, and BRCT domain containing proteins implicated in DNA damage responses. This cluster suggests a role for RNA processing in the M1/G transition. Analysis of genes in clusters for

CURRENT PERSPECTIVE ESSAY

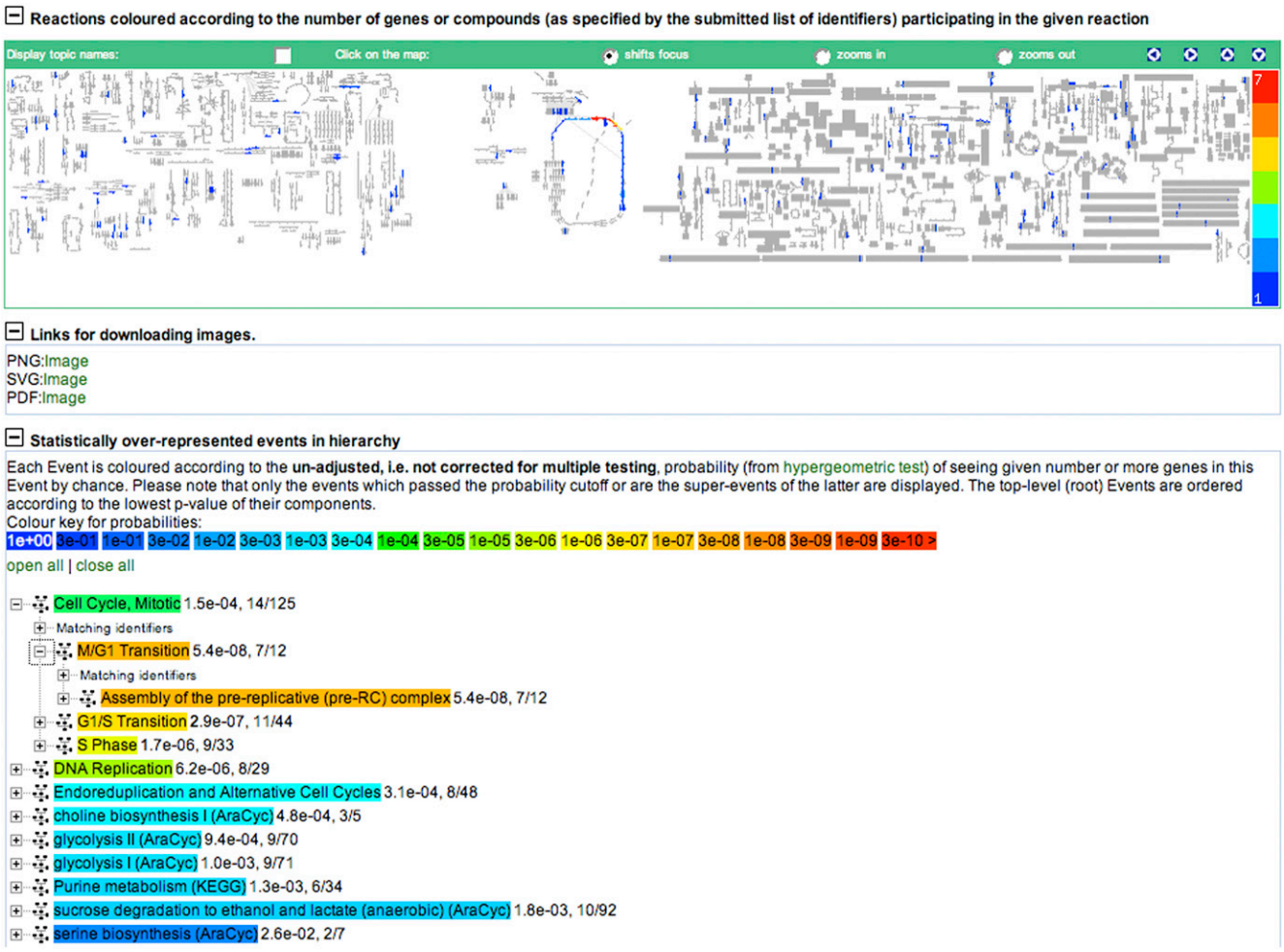


Figure 3. SkyPainter Overrepresentation Analysis of Glucose-Responsive Genes.

A total of 369 genes were maximally induced by glucose at 6 h (fold change >2.5 compared with control at 0 h). SkyPainter identified 79 matching genes within 87 events. The output is shown in two formats followed by an ordered list of pathways showing statistically overrepresented events. At the top is the reaction map showing the reactions colored according to the number of submitted genes participating in a given reaction. Underneath is a list of pathways colored according to the probability (using the hypogeometric test) of the number of genes participating in a given pathway by chance. A low probability shows the genes to be statisically overrepresented in the pathway.

coexpression using Prime (http://prime.psc.riken.jp) (see Supplemental Methods online) showed that genes in Cluster 2 are all significantly coregulated, further suggesting that they perform a common function (see Supplemental Table 2 online). Cluster 5 is based on the Arabidopsis Reactome entity ATEB1c that locates to the mitotic spindle during spindle assembly then relocates to the phragmoplast during cytokinesis and the Reactome entity 26S proteasome subunit interacting with CDKA:1. Cluster 5 links several other 26S proteasome subunits, including the regulatory subunits RPN3, 5, and 6, RPT4A, a histone variant, and the MAD2 mitotic spindle checkpoint protein. This identifies the potential involvement of a proteasome complex at this stage of mitosis. Later during chromosome segregation the Arabidopsis Reactome entity Ccs52, a WD40 protein interacts with the anaphase promoting complex. This protein is in Cluster 7, which contains two other WD40 proteins and two CDC20 paralogs (also Arabidopsis Reactome entities) that inactivate the anaphase promoting complex. Thus, Arabidopsis Reactome integrates two protein clusters into the regulation of late-phase mitosis. Cluster 6 contains multiple Arabidopsis Reactome entities involved in the G2 phase and the G2/M transition. Several importin α-subunits mediating nuclear transport are in this cluster. One of these, IMPA-2, interacts with CDKD;2. Cluster 6 also identifies an interaction between the 26

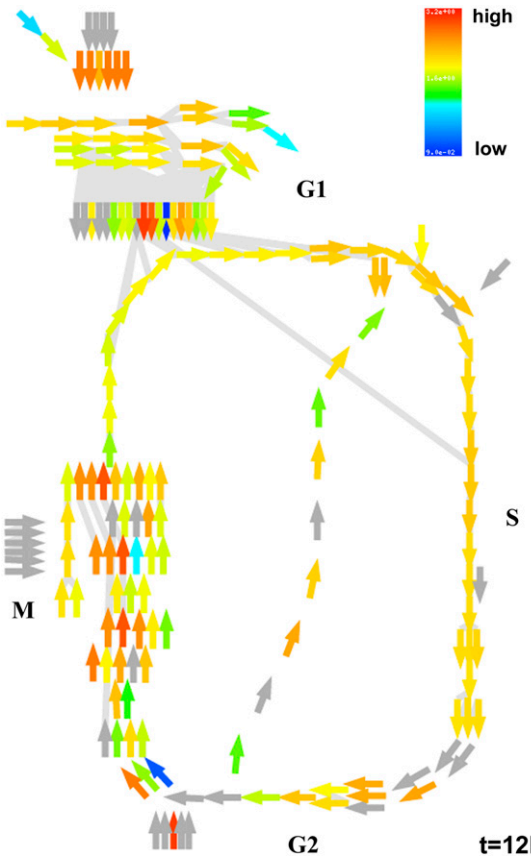


Figure 4. SkyPainter Analysis of a Cell Cycle Time-Course Experiment.

Gene expression was measured at 10 time points over 19 h in *Arabidopsis* cells synchronized for cell cycle progression. SkyPainter colors the reaction arrows according to the average of the expression values (log10 signal) of all components in the reaction. Expression at the 12-h time point shows the reactions within mitosis reaching peak expression; a movie of gene expression changes can be viewed in Supplemental Movie 1 online.

proteasome subunit RPN1 with CDKA;1 at G1 and G2 phases. These analyses demonstrate that the curated pathways in Arabidopsis Reactome provide a comprehensive bioinformatics platform for structuring and integrating the increasingly complex and voluminous data related to cell cycle control and identifies candidate proteins for new experiments.

DATA AND SOFTWARE AVAILABILITY

The data contained in Arabidopsis Reactome can be viewed in Cytoscape (Shannon et al., 2003) and Protégé (http://protege.stanford.edu/) or can be exported in SBML level 2 (Finney and Hucka, 2003), BioPax (http://www.biopax.org/), SVG, and PDF format. The entire content of the Arabidopsis Reactome and tools for curating biological pathways can be downloaded from the Arabidopsis Reactome website by following the ''Download'' link.

COMPARISON AND INTEGRATION WITH OTHER PATHWAY RESOURCES

There has been a rapid growth in the availability of pathway tools since the focus of biological research has moved toward a systems understanding of biological processes. The important

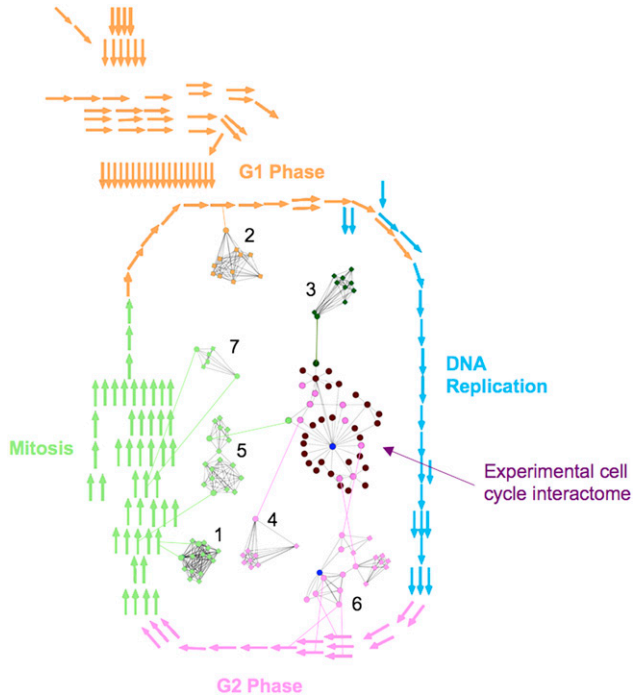


Figure 5. Integrating Experimental and Predicted Protein Interaction Data with the Cell Cycle Network.

The Arabidopsis Reactome cell cycle (shown as the series of arrows colored to represent the different phases of the cell cycle) was merged with the experimental cell cycle protein interactome (shown in the center of the cell cycle). First neighbors from AtPID were also added to the network (represented as squares within the clusters). The MCODE clustering algorithm was then used to predict clusters of proteins that could potentially interact with the cell cycle. The top seven clusters are shown and colored according to the phase of the cell cycle they interact with. The clusters are as follows: Cluster 1, MAP/Shaggy kinase; Cluster 2, WD40 repeat and others; Cluster 3, Aldo/keto reductase; Cluster 4, chaperonin-type protein; Cluster 5, proteasome subunit, prefoldin and microtubule end binding and others; Cluster 6, CDKs, importin, proteasome subunit; Cluster 7, WD40 repeat. CDKA is shown in dark blue in both Cluster 6 and the experimental cell cycle interactome. CDKA plays a role in all phases of the cell cycle. The maroon circles are interacting proteins identified experimentally. Some of the proteins have more than one hit in the Arabidopsis Reactome cell cycle, often because the protein is part of a complex; a link is only made to one reaction for clarity.

CURRENT PERSPECTIVE ESSAY

features of a pathway tool are ease of use, quality of data model and experimental evidence, wide coverage, standard format, facilities to integrate data sets, and that it is open source and open access. AraCyc and KEGG are both comprehensive and standardized pathway resources from which we have already incorporated the data into Arabidopsis Reactome. AraCyc has the widest coverage, mainly of metabolic pathways, many of which have been manually curated. The standard BioPAX exchange format means that these pathways are widely used in other pathway resources, such as MetNet (Wurtele et al., 2007) and Virtual Plant (Gutierrez et al., 2007). KEGG Pathway is limited by the data being mainly homology based. BioPathAt (Lange and Ghassemian, 2005) and MetNet are both sources of curated pathway information that can be used to aid the manual curation of pathways in the future; however, the lack of standardized exchange formats means that these pathways cannot be automatically incorporated into Arabidopsis Reactome. MAPMAN (Thimm et al., 2004) is a user-driven tool for displaying large genomics data sets onto diagrams of metabolic pathways and other processes. Its wide coverage of metabolic and other biological processes, expert curation, and useful hierarchical ontologies with gene mappings make it a popular choice for integrating datasets. However, it lacks a standard format for exchanging pathway information with other pathway resources, and it lacks the level of detail of reactions compared with Reactome due to its ontology-based structure. Currently, we are developing a common exchange format between Arabidopsis Reactome and MAPMAN.

SUMMARY AND FUTURE PROSPECTS

The human Reactome team has established a data model and software tools that allow scientists to capture a wide variety of biological knowledge in a form allowing computational manipulation, and these have been developed as a knowledgebase for human biological processes (Joshi-Tope et al., 2005). Apart from the human Reactome, there are currently four other Reactomes under development for *Drosophila melanogaster*, *Caenorhabditis elegans*, *Archea*, and chicken. In this essay, we have demonstrated the application of the Reactome system to the reference plant species *Arabidopsis* and shown how the data model accurately captures many aspects of different biological processes. Although these models currently encompass only a small fraction of the molecular entities in *Arabidopsis*, this is an important first step toward a quality-assured whole-genome network reconstruction.

The current release of Arabidopsis Reactome (version 2) describes the functions of 276 proteins (233 reactions) curated from 262 primary literature sources. These proteins function in a number of pathways, including the mitotic cell cycle, DNA synthesis, cell cycle checkpoints, alternative cell cycles (including endoreduplication), and several metabolic pathways. Additionally, 1919 proteins have been imported from KEGG (Kanehisa and Goto, 2000) and AraCyc (Mueller et al., 2003) pathway databases. Together with the curated proteins, they represent 8% of the estimated 27,238 *Arabidopsis* protein-coding genes (http://www.arabidopsis.org/portals/genAnnotation/). The curated and imported proteins in Arabidopsis Reactome appear in a total of 8269 reactions from 318 superpathways. The *Arabidopsis* proteome was electronically projected in Arabidopsis Reactome onto five other published plant and moss predicted proteomes via orthologous transfer. This allows the visualization of the equivalent reactions and pathways on other sequenced plant species and assesses pathway conservation between species. Researchers can interact with the web-based interface to browse pathways and visualize ''-omics'' data on the reaction map.

Arabidopsis Reactome pathways allow users to formulate hypotheses based on a much wider range of data and knowledge than hitherto possible. These include studies of network topology (Siegal et al., 2007), network properties such as instability (Wilhelm, 2007) the determination of system attractors (Bornholdt, 2005; Davidich and Bornholdt, 2008), metabolic flux analysis (Morgan and Rhodes, 2002), and the ability to validate algorithms used for the reverse engineering of networks from data (Bansal et al., 2007). The addition of kinetic functions and their parameters will enable approaches to dynamical modeling and metabolic control analysis (Morgan and Rhodes, 2002). We have demonstrated three examples of this functionality. First, orthologous transfer from *Arabidopsis* leads to the identification of candidate genes that are potentially involved in the same pathway in other species and identified potential gaps in gene function between plant species. Such predictions can be tested using functional genomics approaches and reannotation. Second, overlaying gene expression data on pathways using SkyPainter identified coordinately regulated genes within pathways and visualizes gene expression data in a clear and intuitive way. Finally, by merging curated Arabidopsis Reactome modules, such as the cell cycle with experimental (Van Leene et al., 2007) and predicted protein–protein interactions (Cui et al., 2007), we identified clusters of genes that can be tested for cell cycle–related phenotypes. It is conceivable that any other pairwise inference (e.g., genetic interaction or protein domain) could also be used as a connection (edge or weighted edge) in the Cytoscape graph and analyzed with a network query application, such as MCODE, for example, to establish genetic interaction networks.

Most of the knowledge of biological processes in plants is described as free text in the published literature and other resources. The Reactome approach is to extract this distributed knowledge using authors and curators who understand the scientific field and can make expert judgments on what constitutes good evidence and can resolve ambiguities in the use of natural language. This process requires a high degree of knowledge and commitment, but once achieved for reference species such as *Arabidopsis*, it will potentially benefit the wider plant research community since such pathway knowledge can be projected electronically onto other plant genomes using Reactome. Currently, knowledge from *Arabidopsis* has been

electronically applied to poplar, rice, grape, and *P. patens* and is available from the Arabidopsis Reactome website. With increased content, it should be possible to annotate newly sequenced plant genomes by orthologous transfer from Arabidopsis Reactome using algorithms such as OrthoMCL (Li et al., 2003) and InParanoid (Remm et al., 2001). As 35 plant genome-sequencing projects are currently underway (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi?p3=11:Plants), this function of Arabidopsis Reactome will be highly relevant.

Current methods of knowledge generation involve the production of text and images that are largely intractable to electronic manipulation. Text mining is currently not capable of reconstructing this knowledge in a computational form (Jensen et al., 2006); therefore, it is currently merely an aid to the curation process rather than a comprehensive solution. One way to facilitate the growth and uptake of Arabidopsis Reactome among users is through a collaborative model involving the editorial process and publishers. Such an approach has recently been proposed that will require authors to deposit data in the TAIR database as part of the publication process (http://www.arabidopsis.org/news/plant_phys_partnership.txt). Arabidopsis Reactome is ideally suited for this type of activity as it captures a wide variety of knowledge about biological entities, such as genes, transcripts, proteins, modified proteins, and metabolites. This and other examples would promote its use, and the benefits of high-quality electronically available knowledge would be realized more widely. The standardization of pathway knowledge representation is critically important in pathway curation as it aids the integration of data in systems biology to generate new hypotheses (Draghici et al., 2007). Arabidopsis Reactome provides a powerful foundation, based on careful and comprehensive curation of the literature, which will facilitate many areas of biological research in plants and help to integrate biological knowledge.

Supplemental Data

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

Supplemental Figure 1. SkyPainter Overrepresentation Analysis of Glucose Upregulated Genes.

Supplemental Figure 2. Arabidopsis Reactome Cell Cycle Merged with the Experimental Cell Cycle Interactome Visualized in Cytoscape.

Supplemental Figure 3. Cell Cycle Network Merged with AtPID Protein–Protein Interaction Data.

Supplemental Table 1. Description of Predicted Protein Clusters 1 to 7 Using MCODE and Their Location in the Cell Cycle Network.

Supplemental Table 2. Coexpression Analysis of Genes within Each Protein–Protein Interaction Cluster Using the Correlated Gene Data from the PRIME Database.

Supplemental Methods.

Supplemental Movie 1. SkyPainter Movie Showing Gene Expression Over 10 Time Points during the Cell Cycle.

ACKNOWLEDGMENTS

We thank the Reactome development team for their support and useful criticism and Yunhai Li for the glucose microarray data and analysis. This work was supported by funds from the Biotechnology and Biological Sciences Research Council (BBSRC) under the BBSRC Bioinformatics and e-Science initiative (BBS/B/13829) and by the European Commission Project Arabidopsis GROwth Network integrating OMICS technologies (Contract 037704). E.E.S., I.V., E.B., G.W., P.D., and L.D.S. are supported by a grant from the U.S. National Institutes of Health, a grant from the European Union Sixth Framework Programme, and subcontracts from the National Institutes of Health Cell Migration Consortium and the European Bioinformatics Institute Industry Programme.

REFERENCES

Alonso, J.M., and Ecker, J.R. (2006). Moving forward in reverse: Genetic technologies to enable genome-wide phenomic screens in *Arabidopsis.* Nat. Rev. Genet. **7:** 524–536.

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

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

Bader, G.D., and Hogue, C.W. (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics **4:** 2.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., and di Bernardo, D. (2007). How to infer gene networks from expression profiles. Mol. Syst. Biol. **3:** 78.

Bornholdt, S. (2005). Systems biology. Less is more in modeling large genetic networks. Science **310:** 449–451.

Clark, R.M., et al. (2007). Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana.* Science **317:** 338–342.

Craigon, D.J., James, N., Okyere, J., Higgins, J., Jotham, J., and May, S. (2004). NASCArrays: A repository for microarray data generated by NASC's transcriptomics service. Nucleic Acids Res. **32:** D575–D577.

Cui, J., Li, P., Li, G., Xu, F., Zhao, C., Li, Y., Yang, Z., Wang, G., Yu, Q., Li, Y., and Shi, T. (2007). AtPID: *Arabidopsis thaliana* protein interactome database—An integrative platform for plant systems biology. Nucleic Acids Res. **36:** D999–D1008.

Davidich, M.I., and Bornholdt, S. (2008). Boolean network model predicts cell cycle sequence of fission yeast. PLoS ONE **3:** e1672.

de Bono, B., Vastrik, I., D'Eustachio, P., Schmidt, E., Gopinath, G., Croft, D., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., Wu, G., Birney, E., and Stein, L. (2007). Reactome: An integrated expert model of human molecular processes and access toolkit. J. Integr. Bioinformatics **4:** 84.

Degtyarenko, K., Matos, P.D., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcantara, R., Darsow, M., Guedj, M., and Ashburner, M. (2007). ChEBI: A database and ontology for chemical entities of biological interest. Nucleic Acids Res. **36:** D344–D350.

Draghici, S., Khatri, P., Tarca, A.L., Amin, K., Done, A., Voichita, C., Georgescu, C., and Romero, R. (2007). A systems biology approach for pathway level analysis. Genome Res. **17:** 1537–1545.

Feller, W. (1968). An Introduction to Probability Theory and Its Applications. (New York: John Wiley & Sons).

Finney, A., and Hucka, M. (2003). Systems biology markup language: Level 2 and beyond. Biochem. Soc. Trans. **31:** 1472–1473.

CURRENT PERSPECTIVE ESSAY

Furukawa, T., Ishibashi, T., Kimura, S., Tanaka, H., Hashimoto, J., and Sakaguchi, K. (2003). Characterization of all the subunits of replication factor C from a higher plant, rice (*Oryza sativa* L.), and their relation to development. Plant Mol. Biol. **53:** 15–25.

Geisler-Lee, J., O'Toole, N., Ammar, R., Provart, N.J., Millar, A.H., and Geisler, M. (2007). A predicted interactome for Arabidopsis. Plant Physiol. **145:** 317–329.

Gutierrez, R.A., Lejay, L.V., Dean, A., Chiaromonte, F., Shasha, D.E., and Coruzzi, G.M. (2007). Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in *Arabidopsis.* Genome Biol. **8:** R7.

Heazlewood, J.L., Verboom, R.E., Tonti-Filippini, J., Small, I., and Millar, A.H. (2007). SUBA: The Arabidopsis subcellular database. Nucleic Acids Res. **35:** D213–D218.

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

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

Jensen, L.J., Saric, J., and Bork, P. (2006). Literature mining for the biologist: From information retrieval to biological discovery. Nat. Rev. Genet. **7:** 119–129.

Joshi-Tope, G., et al. (2005). Reactome: A knowledgebase of biological pathways. Nucleic Acids Res. **33:** D428–D432.

Joshi-Tope, G., Vastrik, I., Gopinath, G.R., Matthews, L., Schmidt, E., Gillespie, M., D'Eustachio, P., Jassal, B., Lewis, S., Wu, G., Birney, E., and Stein, L. (2003). The Genome Knowledgebase: A resource for biologists and bioinformaticists. Cold Spring Harb. Symp. Quant. Biol. **68:** 237–243.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. **28:** 27–30.

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. (2002). The KEGG databases at GenomeNet. Nucleic Acids Res. **30:** 42–46.

Kohler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Ruegg, A., Rawlings, C., Verrier, P., and Philippi, S. (2006). Graph-based analysis and visualization of experimental results with ONDEX. Bioinformatics **22:** 1383–1390.

Lange, B.M., and Ghassemian, M. (2005). Comprehensive post-genomic data analysis approaches integrating biochemical pathway maps. Phytochemistry **66:** 413–451.

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

Li, Y., Lee, K.K., Walsh, S., Smith, C., Hadingham, S., Sorefan, K., Cawley, G., and Bevan, M.W. (2006). Establishing glucose- and ABA-regulated transcription networks in *Arabidopsis* by microarray analysis and promoter classification using a Relevance Vector Machine. Genome Res. **16:** 414–427.

Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. (2005). Entrez Gene: Gene-centered information at NCBI. Nucleic Acids Res. **33:** D54–D58.

Menges, M., Hennig, L., Gruissem, W., and Murray, J.A. (2003). Genome-wide gene expression in an *Arabidopsis* cell suspension. Plant Mol. Biol. **53:** 423–442.

Ming, R., et al. (2008). The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature **452:** 991–996.

Morgan, J.A., and Rhodes, D. (2002). Mathematical modeling of plant metabolic pathways. Metab. Eng. **4:** 80–89.

Mueller, L.A., Zhang, P., and Rhee, S.Y. (2003). AraCyc: A biochemical pathway database for Arabidopsis. Plant Physiol. **132:** 453–460.

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

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

Rhee, S.Y., et al. (2003). The *Arabidopsis* Information Resource (TAIR): A model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. Nucleic Acids Res. **31:** 224–228.

Schneider, M., Bairoch, A., Wu, C.H., and Apweiler, R. (2005). Plant protein annotation in the UniProt Knowledgebase. Plant Physiol. **138:** 59–66.

Schoof, H., Ernst, R., Nazarov, V., Pfeifer, L., Mewes, H.W., and Mayer, K.F. (2004). MIPS *Arabidopsis thaliana* Database (MAtDB): An integrated biological knowledge resource for plant genomics. Nucleic Acids Res. **32:** D373–D376.

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

Shultz, R.W., Tatineni, V.M., Hanley-Bowdoin, L., and Thompson, W.F. (2007). Genome-wide analysis of the core DNA replication machinery in the higher plants Arabidopsis and rice. Plant Physiol. **144:** 1697–1714.

Siegal, M.L., Promislow, D.E., and Bergman, A. (2007). Functional and evolutionary inference in gene networks: Does topology matter? Genetica **129:** 83–103.

Smertenko, A.P., Chang, H.Y., Sonobe, S., Fenyk, S.I., Weingartner, M., Bogre, L., and Hussey, P.J. (2006). Control of the AtMAP65-1 interaction with microtubules through the cell cycle. J. Cell Sci. **119:** 3227–3237.

Somerville, C., and Koornneef, M. (2002). A fortunate choice: The history of *Arabidopsis* as a model plant. Nat. Rev. Genet. **3:** 883–889.

Suderman, M., and Hallett, M. (2007). Tools for visually exploring biological networks. Bioinformatics **23:** 2651–2659.

Thimm, O., Bläsing, O., Gibon, Y., Nagel, A., Meyer, S., Krüger, P., Selbig, J., Müller, L.A., Rhee, S.Y., and Stitt, M. (2004). MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. **37:** 914–939.

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

Van Leene, J., et al. (2007). A tandem affinity purification-based technology platform to study the cell cycle interactome in *Arabidopsis thaliana.* Mol. Cell. Proteomics **6:** 1226–1238.

Vastrik, I., et al. (2007). Reactome: A knowledge base of biologic pathways and processes. Genome Biol. **8:** R39.

Velasco, R., et al. (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS ONE **2:** e1326.

Wilhelm, T. (2007). Analysis of structures causing instabilities. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. **76:** 011911.

Wurtele, E.S., et al. (2007). MetNet: Systems biology software for Arabidopsis. In Concepts in Plant Metabolomics, B.J. Nikalau and E.S. Wurtele, eds (Dordrecht, The Netherlands: Springer), pp. 145–158.

Zimmermann, P., Hennig, L., and Gruissem, W. (2005). Gene-expression analysis and network discovery using Genevestigator. Trends Plant Sci. **10:** 407–409.

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., and Gruissem, W. (2004). GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox. Plant Physiol. **136:** 2621–2632.