# VirtualPlant: A Software Platform to Support Systems Biology Research[1][W][OA]

**Manpreet S. Katari[2], Steve D. Nowicki[2], Felipe F. Aceituno, Damion Nero, Jonathan Kelfer, Lee Parnell Thompson, Juan M. Cabello, Rebecca S. Davidson, Arthur P. Goldberg, Dennis E. Shasha, Gloria M. Coruzzi, and Rodrigo A. Gutiérrez***

Center for Genomics and Systems Biology, Department of Biology (M.S.K., S.D.N., D.N., J.K., L.P.T., R.S.D., A.P.G., G.M.C., R.A.G.), and Courant Institute of Mathematical Sciences (D.E.S.), New York University, New York, New York 10003; and Departmento de Genética Molecular y Microbiología, P. Universidad Católica de Chile, Casilla 114–D, Santiago, Chile (F.F.A., J.M.C., R.A.G.)

Data generation is no longer the limiting factor in advancing biological research. In addition, data integration, analysis, and interpretation have become key bottlenecks and challenges that biologists conducting genomic research face daily. To enable biologists to derive testable hypotheses from the increasing amount of genomic data, we have developed the VirtualPlant software platform. VirtualPlant enables scientists to visualize, integrate, and analyze genomic data from a systems biology perspective. VirtualPlant integrates genome-wide data concerning the known and predicted relationships among genes, proteins, and molecules, as well as genome-scale experimental measurements. VirtualPlant also provides visualization techniques that render multivariate information in visual formats that facilitate the extraction of biological concepts. Importantly, VirtualPlant helps biologists who are not trained in computer science to mine lists of genes, microarray experiments, and gene networks to address questions in plant biology, such as: What are the molecular mechanisms by which internal or external perturbations affect processes controlling growth and development? We illustrate the use of VirtualPlant with three case studies, ranging from querying a gene of interest to the identification of gene networks and regulatory hubs that control seed development. Whereas the VirtualPlant software was developed to mine Arabidopsis (*Arabidopsis thaliana*) genomic data, its data structures, algorithms, and visualization tools are designed in a species-independent way. VirtualPlant is freely available at www.virtualplant.org.

Today, experimental biology laboratories usually investigate the molecular mechanisms underlying a physiological or developmental response by identifying the genes involved using a genomic platform, such as microarray (or, soon, deep sequencing) technology. Such a platform might identify genes regulated during a physiological or developmental response. Once the relevant gene sets are identified, biologists next analyze their functional relationships (e.g. whether they belong to the same metabolic pathway) and analyze their properties in the context of known biological pathways (DeRisi et al., 1997). Performing these tasks can be cumbersome because the biologist has to use several different tools to accomplish them. In addition, the difficulty is often increased because the different tools do not read and write the same data formats, forcing the biologist to obtain data conversion software.

Aside from the challenge of integrating the vast amount of knowledge accumulated in the literature about the relevant genes, the genomic data available in the public domain have been obtained with a large number of experimental approaches and an even larger number of laboratories. Moreover, the information is stored in numerous databases, and it is encoded in diverse formats and database schemas. Bioinformatics faces a major challenge integrating this large-scale, heterogeneous information into architectures that support biological research. Different approaches that have been employed include hypertext navigation on the World Wide Web, data warehousing, and client-side integration (for example, see Ritter, 1994; Karp, 1996; Siepel et al., 2001; Philippi, 2004; Wilkinson et al., 2005). Once data from distinct database sources are coherently integrated, tools and computer models can be used to enable one to visualize and analyze this biological data from a systems perspective (Ideker et al., 2001). Several environments have been developed to support data integration and modeling

[1] This work was supported by the National Science Foundation (grant nos. DBI 0445666 to R.A.G., D.E.S., and G.M.C., IOB 0519985 to G.M.C. and D.E.S., and MCB–0209754 to D.E.S.), FONDECYT (grant no. 1060457), Grape Genomics (grant no. CORFO07Genoma01 to R.A.G.), Millennium Nucleus for Plant Functional Genomics (grant no. P06–009–F to R.A.G.), and the National Institutes of Health (grant nos. R01 GM 032877 to G.M.C. and 5F32GM75600 to M.S.K.).

[2] These authors contributed equally to the article.

* Corresponding author; e-mail rgutierrez@bio.puc.cl.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Rodrigo A. Gutiérrez (rgutierrez@bio.puc.cl).

[W] The online version of this article contains Web-only data.

[OA] Open Access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.109.147025

(Kahlem and Birney, 2007). Some software allows detailed mathematical representation of cellular processes (e.g. Gepasi [Mendes, 1997] and Virtual Cell [Loew and Schaff, 2001]), while other software permits qualitative representations of cellular components and their interactions (e.g. Cytoscape [Shannon et al., 2003], Osprey [Breitkreutz et al., 2003], and N-Browse [Kao and Gunsalus, 2008]). Generally, quantitative models build detailed mathematical abstractions of specific cellular process. Quantitative models are powerful because they describe a system in detail (Endy and Brent, 2001), but they require a detailed understanding of the system. Unfortunately, this information is available for only a few biological processes. In fact, there are still many gaps in our qualitative understanding of biological systems, even for model organisms. For example, most of the genes in Arabidopsis (*Arabidopsis thaliana*) have not yet been experimentally characterized. Thus, while quantitative computer models can provide powerful, detailed representations of biological systems, not enough is known about Arabidopsis and other plants to construct such models of them or their major components. Therefore, we have focused on building software that facilitates analysis of the systems and statistical and interaction relationships between their genes and gene products.

Today's most widely available measure of gene function is the level of gene expression provided by a microarray analysis. Many approaches and tools support analysis of expression data. A now classic approach, for example, is to identify genes that are coregulated in their expression patterns across selected experimental conditions (e.g. Eisen et al., 1998). An extensive review of the different software tools that are available for studying gene coexpression is available (Usadel et al., 2009). To identify genes that are differentially expressed between two experimental conditions, statistical methods such as Rank Products can be used (Breitling et al., 2004; Hong et al., 2006). Several tools are available as packages in BioConductor, a project largely composed of tools written in the statistical language R (Gentleman et al., 2004). To determine the biological significance of differentially or coexpressed genes, biologists often evaluate the frequency of occurrence of functional attributes provided by structured functional annotations, such as Gene Ontology (GO; Ashburner et al., 2000). Several software packages to automate this type of analysis now exist (e.g. Onto-Express [Khatri et al., 2002], GoMiner [Zeeberg et al., 2003], GOSurfer [Zhong et al., 2004], and FatiGO [Al-Shahrour et al., 2004]). While advanced data analysis tools for exploiting genomic data are rapidly emerging (for review, see Brady and Provart, 2009), the narrow specialization of most current software tools forces geneticists to employ many tools to analyze the data in a single biological study. This cumbersome and inefficient process greatly hinders biologists following a systems approach of iterative in silico exploration and experimentation.

VirtualPlant addresses these problems by integrating selected genomic data and analysis tools into a single Web-accessible software platform. The goal of our work is to help biologists discover new insights by synthesizing multiple data sources. VirtualPlant provides access to a database storing selected information about Arabidopsis and rice (*Oryza sativa*) experiments, genes, gene products, and their properties. VirtualPlant's software architecture and data model have been designed and created in a generic, species-independent manner to ease the addition of new organisms and tools in the future. The VirtualPlant database also includes a high-level representation of plant cellular components and interactions that allow users to create molecular networks "on the fly." These molecular networks provide a framework for analyzing experimental measurements. VirtualPlant also includes novel data visualization and data analysis techniques that allow seamless information exploration across many data sets with the help of a shopping cart in which gene sets from experiments and/or analyses can be stored and then used as inputs to other tools to enable iterative analysis. For concreteness, we present an example of how we have used VirtualPlant to identify gene networks and putative regulatory hubs that control seed development. We have previously demonstrated the use of VirtualPlant and specific tools embodied in the VirtualPlant system to generate hypotheses that were validated experimentally (Wang et al., 2004; Gutiérrez et al., 2007b, 2008; Gifford et al., 2008; Thum et al., 2008).

## RESULTS

### The VirtualPlant Data and Tools

VirtualPlant was constructed on top of a small data warehouse that supports the data analysis process. This warehouse includes descriptions of molecular entities (e.g. gene annotations and functional classification), molecular interactions (metabolic associations, regulatory interactions, and other interaction data from public databases), and publicly available microarray data (including more than 1,800 gene chip hybridizations from the ATH1 Affymetrix platform obtained from the European Arabidopsis Stock Center [NASC] using the Affywatch subscription service). A description of the currently supported data types and corresponding sources can be found in Table I. VirtualPlant contains a software module that automatically refreshes this database on a regular basis. VirtualPlant's interface was designed to be analogous to the familiar E-commerce paradigm, which has customers (aka biologists) and inventory (aka data; Fig. 1). Users can interact with data in VirtualPlant in three main ways: (1) browse the database,(2) query the database content, and (3) upload their own data. The VirtualPlant Web site is divided into four separate windows: (1) the navigation window located on the top, (2) the

**Table I.** *Data available in the VirtualPlant database*

| Data | Source | Statistics | Reference |
|---|---|---|---|
| Gene annotation | TAIR | 33,264 genes | Rhee et al. (2003) |
| Functional categories | GeneOntology (TAIR) | 102,879 associations | Ashburner et al. (2000) |
| | MIPSFuncat (MIPS) | 46,514 associations | Mewes et al. (2004) |
| Microarray data | Data files (NASC) | 499 experiments containing 3,829 hybridizations | Craigon et al. (2004); Redman et al. (2004) |
| | Probe to gene associations (AFFYMETRIX) | 22,810 probes mapped to 23,334 genes | Rhee et al. (2003) |
| Biochemical pathways | KEGG | 11,197 | Mueller et al. (2003) |
| | ARACYC | 17,498 | Kanehisa et al. (2004) |
| Regulatory interactions | AGRIS | 343 interactions | Davuluri et al. (2003) |
| Predicted regulatory interactions | | 21,698,658 transcription factors to target predictions | Gutiérrez et al. (2008) |
| | INTERACTOME | 39,317 interactions | Geisler-Lee et al. (2007) |
| | AtPID | 24,418 | Cui et al. (2008) |
| | BIND | 949 | Bader et al. (2002) |
| | MADS BOX | 263 | de Folter et al. (2005) |
| | Calmodulin | 755 calmodulins | Popescu et al. (2007) |
| Literature-based interactions | GENEWAYS | 107 interactions | Rzhetsky et al. (2004) |
| MicroRNA:mRNA interactions | Collated by Dr. Pam Green's laboratory (mirBASE and ASRP) | 582 interactions | Gustafson et al. (2005); Lu et al. (2005); Griffiths-Jones et al. (2006) |

cart window located on the left, (3) the data browser window located on the lower left, and (4) the analysis window located in the middle (Fig. 2). The navigation window provides links to the different features of VirtualPlant. The data browser window provides access to some of the different annotations and functional categories that are loaded into the VirtualPlant database. The analysis window is where most of the activity occurs. Figure 2 shows the "analysis" view, which is the result from clicking on "analyze" in the navigation window. The pull-down menu shows the different types of functions and tools that are available for that species (Arabidopsis) and data type.

As discussed above, a key challenge to analyzing genomic data is the complex analysis workflow required by currently available software. VirtualPlant solves this problem by integrating multiple tools into a single platform that standardizes the representation of their inputs and outputs so that the output of almost any analysis can be stored in VirtualPlant and later input to any VirtualPlant analysis tool. These intermediate results are stored indefinitely as sets of genes (or experiments) in the gene cart (Fig. 2). This iterative model enables biologists to make arbitrarily complex, multistep analyses of their genomic data. Furthermore, they can suspend or resume any analysis at any time, returning to VirtualPlant to continue working with previously created intermediate results. In this sense, VirtualPlant is not a single-service site where data are uploaded from a user (biologist), analyzed with a tool, and then downloaded back to the user. Instead, users can iteratively analyze their data by using the output of one data analysis/visualization tool as the input of another tool using the cart as an intermediate. This unique feature of VirtualPlant facilitates a fundamental methodology of systems biology's iterative cycles of data analysis and experimentation (Ideker et al., 2001; Gutiérrez et al., 2005). Three working examples described below illustrate how VirtualPlant can be used to perform iterative data analyses that build and refine testable biological hypotheses.

### Using VirtualPlant to Drive Iterative Cycles of Systems Biology Research

The purpose of the following three case studies is to describe some of the tools available in VirtualPlant and to illustrate the utility of the software in the integration of genomic data to develop testable hypotheses. The first two case studies illustrate some of the basic functions of VirtualPlant. The third case study provides an advanced application of the software. Each case study provides concrete working examples that can help new users learn how to use the software. Links to step-by-step video tutorials for the three case studies are provided on the Web site.

### Case Study 1: Analysis at the Gene Level

This first case study illustrates an analysis of one gene with VirtualPlant. Suppose a biologist poses the following question: What are the biological processes associated with the genes coexpressed with *NIA1*? We start answering this question by searching the database for *NIA1*. A simple way to query the VirtualPlant database is to use the query form, which can be accessed using the query link on the navigation window (Fig. 2). To perform a query, select type "genes," enter *NIA1* in the "keywords" field, and click the

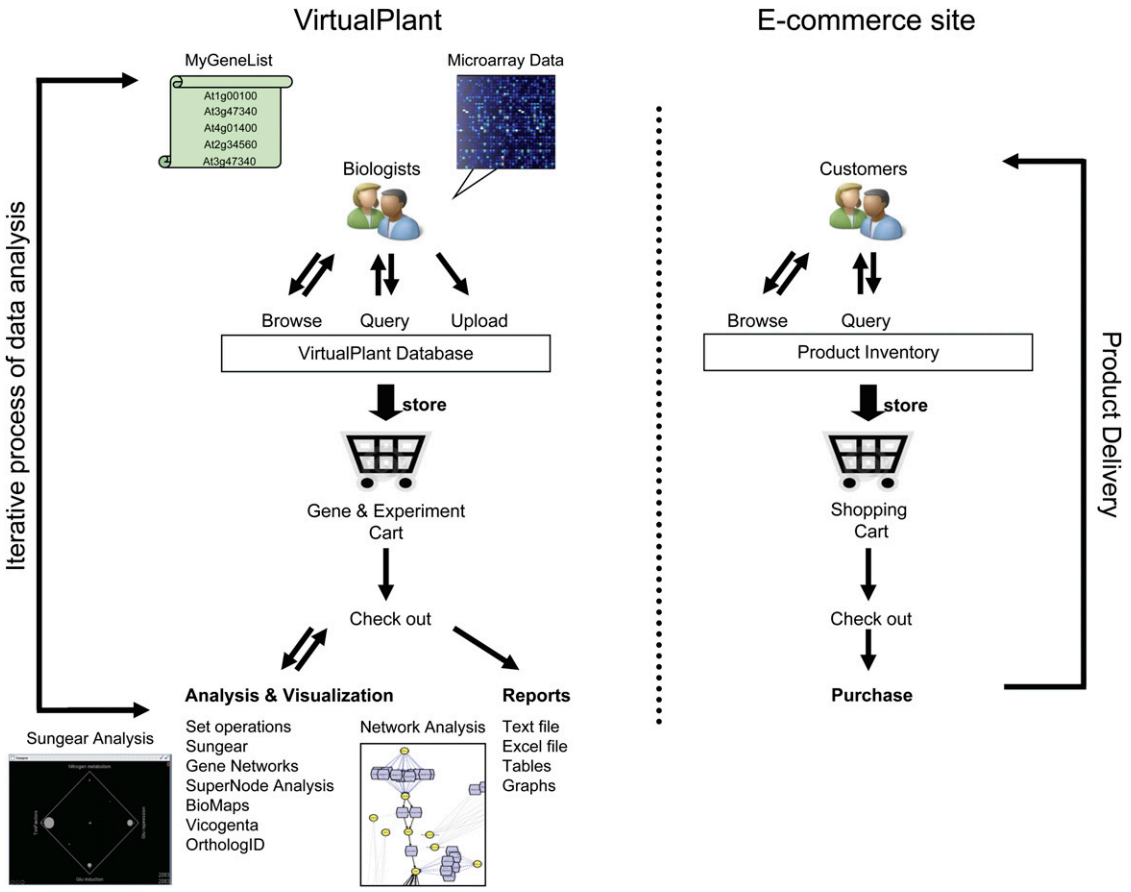



**Figure 1.** Conceptual diagram of the VirtualPlant software system. VirtualPlant follows the e-commerce site logic. In e-commerce sites, users browse and query the database and add products of interest to their shopping cart. Users then check out and purchase the items in their cart. Similarly, VirtualPlant allows biologists to browse lists of genes or microarray experiments with desirable properties. Having found interesting data, they can load the data into the gene cart and "check out" to analyze the selected genes. Biologists can then analyze or visualize the data in the cart to generate biological hypothesis. Most tools in VirtualPlant can store their output in the Cart for a new round of analysis. This key feature allows for iterative filtering and refinement of large data sets.

"submit query" button. The results are displayed in a table where the user can select the result(s) of interest and add it to their cart. Clicking on the gene description "*NIA1,NIA1* (NITRATE REDUCTASE 1)," displays the gene details page, which contains information about the gene, such as its full annotation, gene models, Affymetrix probe ID, and functional annotation terms.

To learn more about the expression of this gene, simply click on the Affymetrix probe ID name (259681_at) that appears in the *NIA1* gene details page. Clicking on 259681_at will open the probe details page. In addition to the probe attributes, this page contains a histogram of the number of probes whose expression correlates with *NIA1* and also displays the correlation values (Fig. 3). These correlation values were determined previously using publicly available microarray experiments from the ATH1 Affymetrix platform obtained from NASC (www.arabidopsis.info). The experiments were first normalized using the RMA method, and all pairwise probe correlations were calculated using Spearman rank correlation. Correlation values between genes can vary based on the experimental data set being examined and the statistics used (Usadel et al., 2009). The purpose of the graph is to show some of the genes that are correlated across a collection of experiments with the gene of interest and then use some of the other tools in the VirtualPlant system to further investigate and explore the coexpressed genes. To select the probes that are correlated to the query probe, one can simply click on the bars of the histogram. The probes that are correlated to the query probe (259681_at in this example) will be displayed in the table under the graph. To select genes that are positively correlated to *NIA1* at a cutoff of at least 0.6 and $<0.7$, one can click on the bar labeled "0.6 to 0.7." This analysis shows that there are 20 probes correlated to 259681_at that map to 23 genes because three of the probes are ambiguous (map to two genes). In order to further analyze the 23 genes

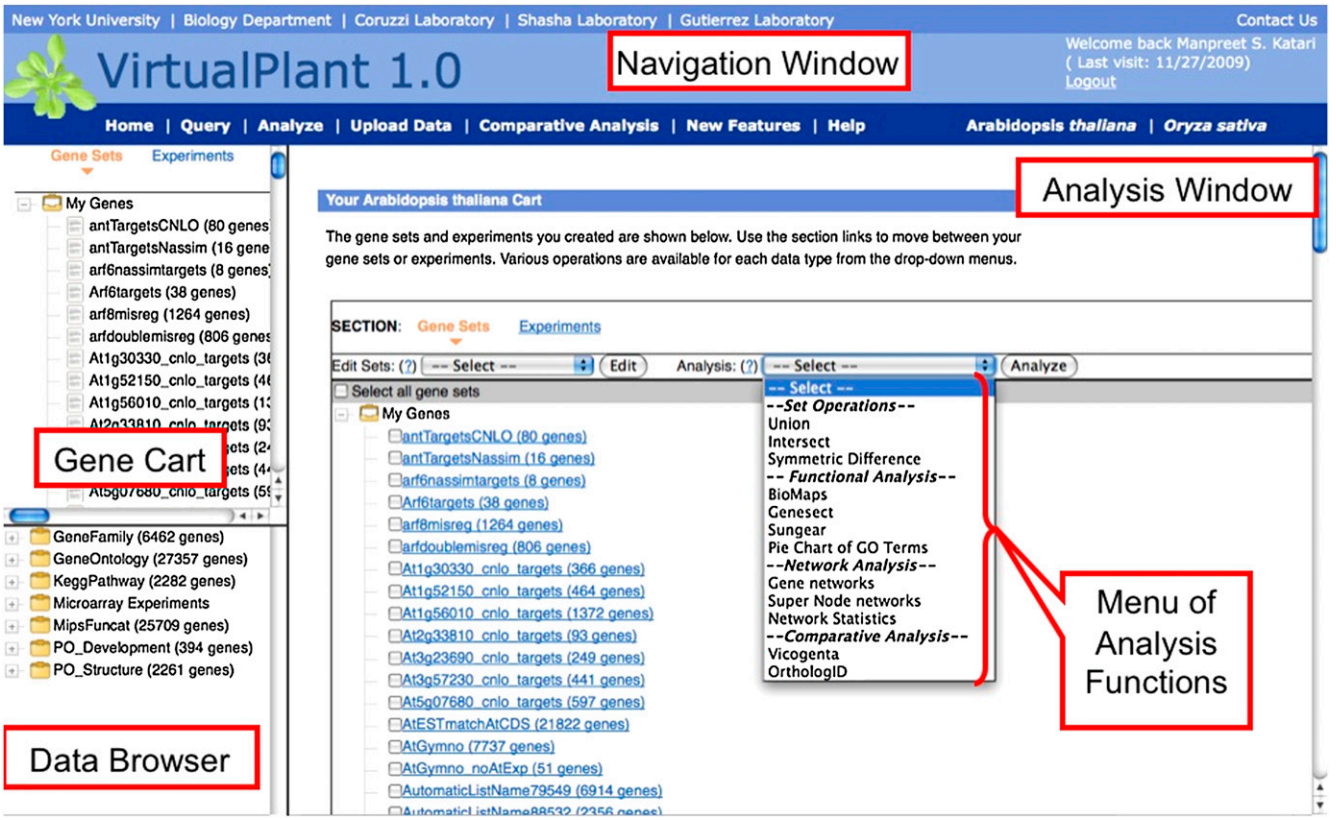



**Figure 2.** The VirtualPlant Web site. There are four main areas in the VirtualPlant Web site: (1) the navigation window (top), (2) the cart window (left), (3) the database browser window (bottom left), and (4) the analysis window (center). The navigation window contains links to the different contents in VirtualPlant. The cart window displays the contents of the cart, which are lists of genes and experiments that have been created and saved by the user. The database browser window allows the user to navigate through different types of data stored in the database. Clicking on "analyze" in the navigation window loads a detailed view of the cart in the analysis window where the user can select the gene or experiment and the different visualization and analysis tools from the pull-down menu.

whose expression correlates with NIA1, one can either export the list of genes to the cart using the "save selection to cart" button or visualize the functional annotation of the genes using the "pie" function (see below). To select all 23 genes, select the first gene, scroll down using the scroll bar on the table, depress the shift key, and select the last gene. Clicking on the "save selection to cart" button creates a new entry named "Corr:259681_at" in the cart. This new list of genes can now be used as input to all other tools in VirtualPlant, such as BioMaps (discussed in case study 2) to find overrepresented GO or Munich Information Center for Protein Sequences (MIPS) terms or gene networks (discussed in case study 3) where one can identify any known or predicted interactions between the 23 genes.

The pie function identifies the biological processes associated with the genes that are correlated to *NIA1* and displays the results in a pie chart. Click on "pie" to open a new window with a pie chart of GO terms (Ashburner et al., 2000) associated with the genes in the selected list (Fig. 3). The pie chart displays the number of genes in each GO term. On the top left there is a pull-down menu where one of the three ontologies (biological process, cellular component, or molecular function) can be selected. By default, all GO terms that are directly associated with the genes are shown, which include GO terms from different levels of the GO hierarchy. Selecting the "level" checkbox allows the user to select a certain depth of the GO hierarchy. At level 1 of biological process, the three most abundant terms are "cellular process," "metabolic process," and "response to stimulus." When you move the slide to level 2, the terms are more specific and more informative.

This simple exercise indicated that expression of the *NIA1* gene correlates with the expression of genes involved in cellular metabolic process, primary metabolic process, biosynthetic process, response to stress, and response to abiotic stimulus (Fig. 3). This result is consistent with our understanding of nitrate reduction and the coordination between this and other metabolic pathways in plants (Sitt et al., 2002). This answer to the original question "What are the biological processes

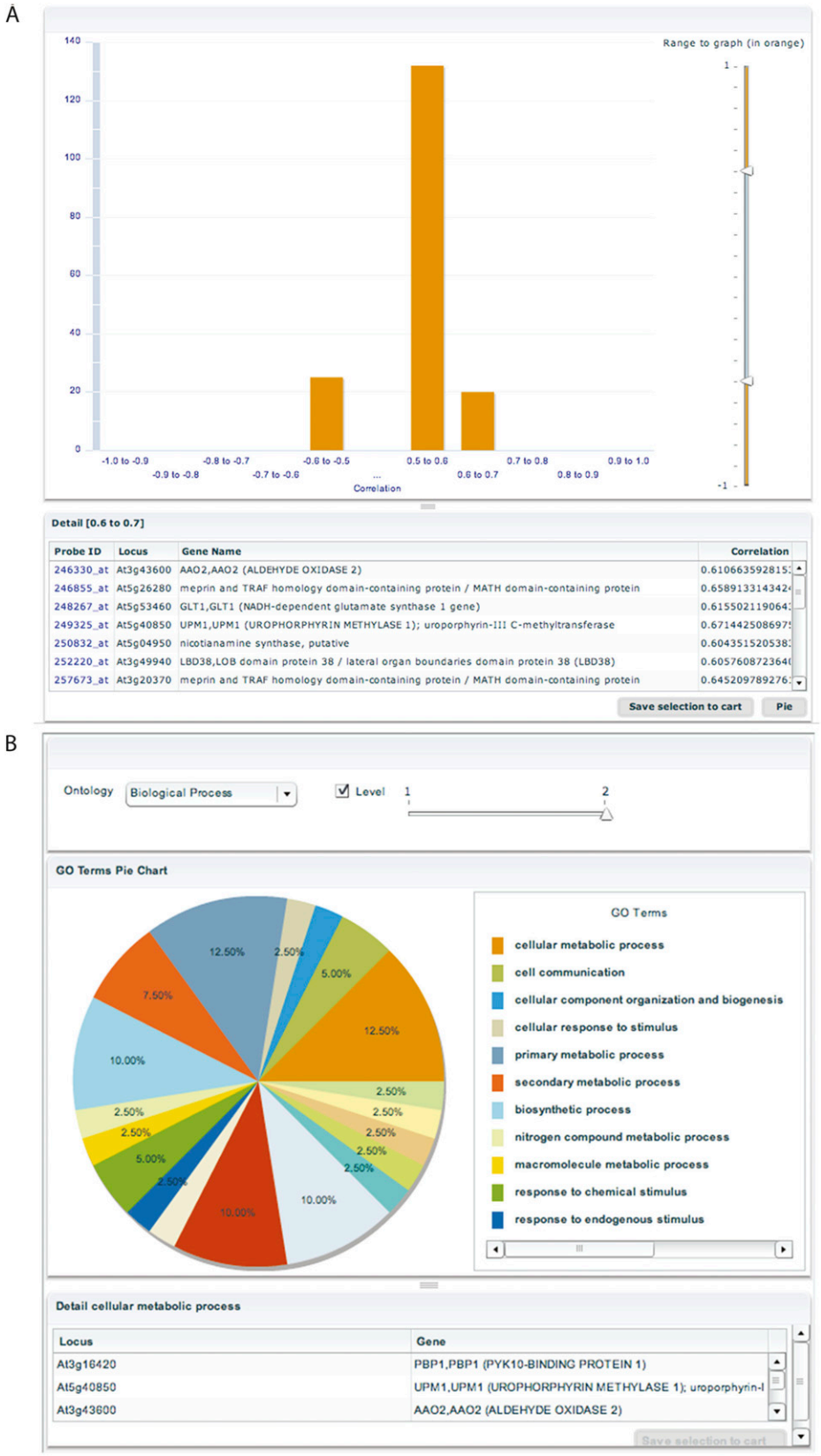



**Figure 3.** Genes correlated to *NIA1* and their gene ontology annotations. A, Histogram representing the number of probes correlated with the NIA1 Affymetrix probe (259681_at). Orange bars represent the number of probes in the different correlation cutoff intervals. These can be selected by the "range to graph" sliding tool on the right of the graph. Clicking on the bars will display probes from the selected interval in the table below. B, Pie-chart of the gene ontology terms associated to 23 genes selected in A. Each term has a different color in the pie chart. The legend to the right of the pie chart indicates the name of the GO term. The pie chart is generated by selecting the genes from the table and clicking on the "pie" button at the bottom.

associated with the genes coexpressed with *NIA1*?" was obtained via VirtualPlant's user-friendly Web interface in a few minutes. The next case study will show how we can use VirtualPlant to obtain statistically significant GO terms associated with a list of genes.

**Case Study 2: Analysis at the Gene List Level**

With the advent of genomic technologies (e.g. microarray technology), many researchers today study not one gene but one or more lists of many genes. These gene lists can be generated in different ways: (1) genes correlated to a gene of interest (previous example), (2) genes in a gene family, (3) genes in a metabolic pathway, or (4) genes that are differentially expressed in several independent microarray experiments. To illustrate how VirtualPlant is used to analyze lists of genes, this case study mines published microarray results of nitrate-regulated genes. Wang et al. (2004) compared global gene expression in response to nitrate treatments in a nitrate reductase (NR)-null mutant and wild-type plants. Genes that are similarly regulated by nitrate in both the wild-type and the NR-null mutant are designated "nitrate-regulated," as the lack of nitrate reductase prevents nitrate reduction and assimilation, thus blocking the production of any downstream metabolic signals. The biological question in this second case study is: "What processes are regulated by nitrate and not a downstream nitrogen signal?" To answer this question, the first step is to identify genes that are regulated similarly by nitrate in both the NR-null mutant (or double mutant) and wild-type plants. To facilitate this demonstration, VirtualPlant provides these lists of genes in the "upload data" page (accessible by clicking on the "upload data" link in the navigation window). The first two lists under "sample data" correspond to genes that are induced in the wild type (439 genes) and induced in the mutant (393 genes). Clicking on the titles of the sample gene lists will add them to the cart. The two sample gene lists will appear in the cart in the top left corner. To find genes that are induced in both mutant and wild-type plants, one uses the "intersect" tool. The intersect tool is available by clicking the "analyze" link in the navigation window (Fig. 2). Once the analysis window has loaded, select the gene sets by clicking the checkboxes in front of the WTRoots and DMRoots lists, choose the "intersect" function from the "analysis" pull-down menu (Fig. 2), and click on the "analyze" button. A new set, which contains genes contained in the two gene lists (wild type and mutant), is created and added to the Cart with the name "Intersection: Wang_etal_2004_I_DMroots/Wang_etal_2004_I_WTroots." In this example, the newly created gene set contains 283 genes that are induced in both the NR-null mutant and in wild-type plants. The three set operation tools (union, intersect, and symmetric difference) input two or more gene lists and produce a gene set output. A highly interactive visual analysis of set operations on two or more lists of genes can also be carried out using Sungear, a tool available from the "analysis" pull-down menu (Fig. 2). For a detailed description of Sungear, please refer to the previous publication by Poultney et al. (2007). An example of the use of Sungear to gain insight into the genomic nitrate response was also published (Gutiérrez et al., 2007a).

To answer the next question, "What processes are regulated by nitrate and not a downstream nitrogen signal?" one needs to identify the biological processes that are significantly overrepresented in the list of 283 genes. One way to answer this question in VirtualPlant is to use the BioMaps tool to determine which GO terms or MIPS functional categories (Mewes et al., 2004) are statistically overrepresented in a list of genes as compared to a background population (e.g. the entire genome). To do this, select the check box to the left of the "Intersection: Wang_etal_2004_I_DMroots/Wang_etal_2004_I_WTroots" list, elect "BioMaps" in the "analysis" pull-down menu (Fig. 2), and click "analyze." Once executed, BioMaps displays a page where the user can select the annotation (GO terms or MIPS), the background population, the statistical method (binomial distribution, hypergeometric distribution, and Fisher's exact test), and the *P* value cutoff to use for the analysis. The *P* values shown in the output of BioMaps are already adjusted for multiple hypotheses testing using false discovery rate correction. For this case study, use the default settings: GO assignments from The Arabidopsis Information Resource (TAIR) and hypergeometric distribution test with a *P* value cutoff of 0.01. Since the gene lists used in this case study were generated using data from ATH1 microarray experiments, we will select the option to use ATH1 genes as the background. The results are provided as "table view," "network view," which is a color-coded graph, and a link to "download to Excel," which is a tab-delimited file that can be opened in Excel, Word, or any other software that can read text files. The final link "unprocessed" downloads a file with comments from the BioMaps analysis. The network view graph is generated by an open source software package called GO::TermFinder (Boyle et al., 2004) and provides an intuitive and visual way to analyze the results. This graph shows the relevant functional terms and their parents as nodes, with annotated genes attached in gray boxes to the most specific term. Clicking on the node name opens up its detail page. The more general terms in the annotation are represented by nodes drawn at the top of the image (e.g. cellular process), with increasing specificity toward the bottom of the image (e.g. cellular carbohydrate metabolic process). The color of the nodes indicates the *P* value of overrepresentation as indicated by the graph's legend (Fig. 4). To simplify the analysis of complex results, this graphical representation of BioMaps will show a maximum of 10 overrepresented functional terms. However, the table output of BioMaps will always contain all significantly

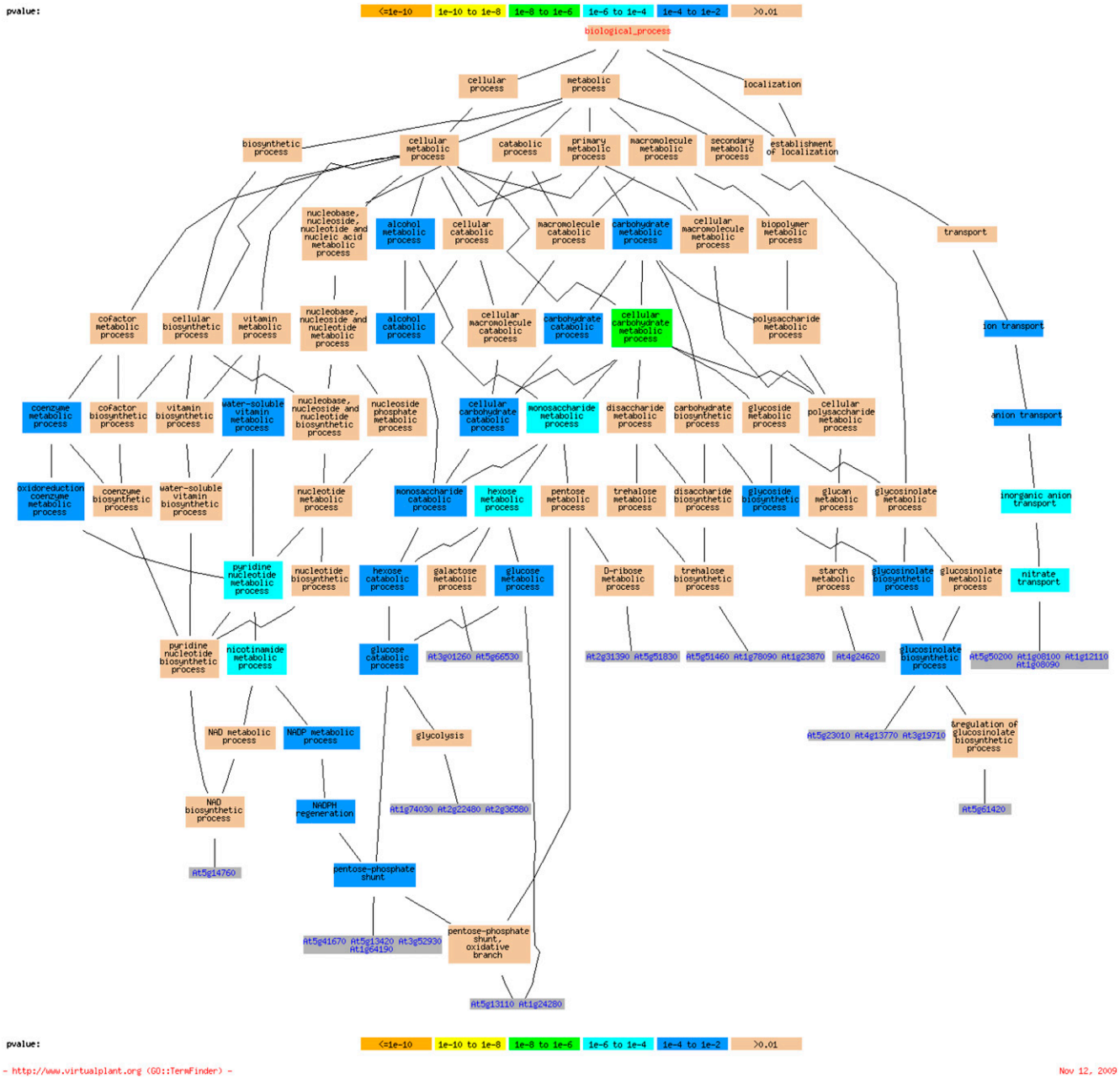



**Figure 4.** BioMaps results of genes that are induced by nitrate in both the wild type and NR-null mutant. BioMaps graphical output is a directed acyclic graph that shows the functional terms that are overrepresented in the gene list analyzed. The gray nodes contain the genes annotated to a functional term. The other colored nodes of the graph correspond to functional terms. The colors indicate the statistical significance of the overrepresentation as indicated in the legend included in the figure. For example, orange nodes correspond to functional terms overrepresented with $P \leq$ 1e-10.

overrepresented terms found by the analysis. The table also provides details regarding the genes in the query list that belong to the term as well as the statistics. Any set of genes listed in the table can be added to the cart. A simple visual inspection of the table view resulting from the BioMaps analysis allows the user to identify the most prominent biological processes in the gene list analyzed (Supplemental Table S1). In this example, cellular carbohydrate metabolic process, alcohol metabolic process, ion transport, and response to abiotic stimulus are some of the overrepresented biological processes among the 283 genes that are regulated by nitrate. This confirms previous results (Crawford, 1995; Sitt et al., 2002; Gutiérrez et al., 2007b) showing that carbohydrate metabolism is a metabolic process that is coordinately regulated with nitrate availability. VirtualPlant's support for gene lists described in this section provides a simple yet powerful way to inte-

grate and analyze published experiments, annotation, pathways, and other data using a list of genes as the common currency. With a few steps, VirtualPlant can help biologists build testable hypotheses from the comparative analysis of genomic data presented in a biological context as shown in a series of recent publications (Gutiérrez et al., 2007b, 2008; Gifford et al., 2008; Thum et al., 2008).

**Case Study 3: Analyzing Gene Networks**

The last case study demonstrates a more advanced use of VirtualPlant. Nitrogen is essential for synthesizing seed storage proteins, which is crucial for proper seed development. The goal of this case study is to determine which nitrogen metabolic genes are controlled at the level of gene expression during seed development and to identify transcription factors that may be key hubs that regulate these genes during seed development. In short, this study asks "What are the regulatory networks responsible for coordinating the expression of genes involved in nitrogen metabolism during seed development?" This case study will demonstrate how to (1) load a publicly available microarray data set into a user's cart, (2) identify genes that are regulated during seed development by determining which genes are differentially expressed, and (3) examine molecular interactions between genes involved in nitrogen metabolism and genes regulated during seed development.

At this point, one must create a VirtualPlant user account, thereby serving two main purposes: (1) allow the user to save gene lists and experiments in the cart and (2) register the user's email so they can be notified when a long-running analysis is completed. For this example, we will start by analyzing microarray data. The microarray experiments can be loaded to the cart either by browsing for the experiment using the data browser window or by uploading the experimental data directly. VirtualPlant accepts two formats for data upload: (1) original ATH1 CEL files, which can then be normalized using either gcRMA or MAS5, and (2) matrices of expression values. The second format allows users to use a different normalization method and then upload the normalized data to VirtualPlant. It also allows users to upload experiments generated with other microarray platforms as well as alternative experimental approaches, such as next-generation sequencing technologies.

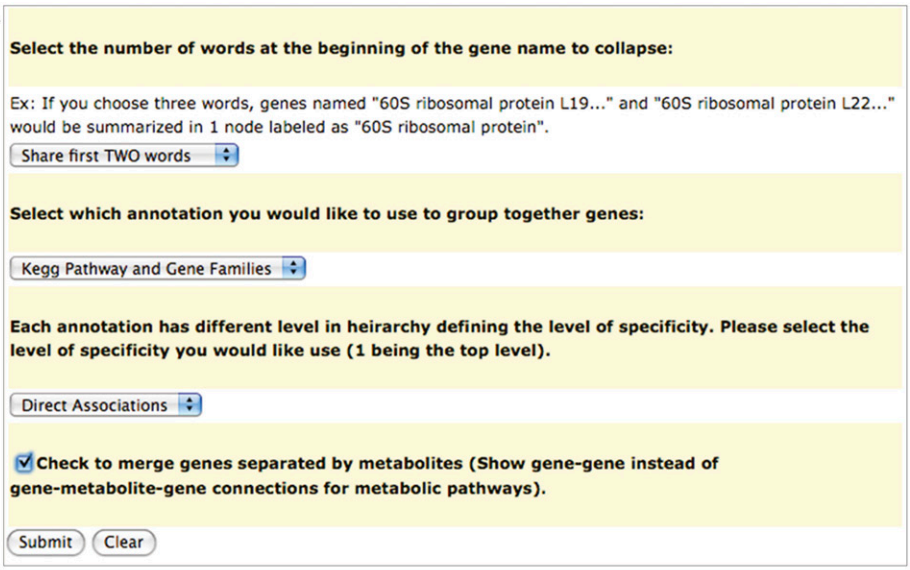
The experiment selected for this case study was a seeds and siliques developmental time series generated by the AtGenExpress project (Schmid et al., 2005). To select this experiment, click on "microarray experiment" in the data browser window at the lower left of the screen. Then, in the main window, click on "AtGenExpress Project," "developmental stage," "developmental series," and then "Detlef Weigel, Jan Lohmann, Markus Schmid AtGenExpress: Developmental series (siliques and seeds) (154)" (Schmid et al., 2005). The user can add this experiment to the cart using the "create experiment" button. To facilitate the demonstration of the VirtualPlant software for first-time users, we also provide a link on the upload data page to directly add this experiment to the cart. To analyze the experiment, click on "analyze" in the navigation window. By default, the analysis view displays the gene sets section of the cart. To view the experiments section, click on the "experiments" link. Now the main window displays all the experiments in the cart. Select the checkbox near the experiment, select "find differentially expressed genes" from the pull-down menu, and click "analyze." A form will appear that allows users to select the "base" and "treatment" microarray hybridizations. Placing the mouse over a slide's name provides more detailed information. In this example, the last five stages of the developmental series are derived from isolated seeds containing no silique tissue, so we will analyze only these last five stages. To identify differentially expressed genes during seed development, select a seed development stage as base and the subsequent stage as treatment. For the first comparison, select all three ATGE_79 as base and all three ATGE_81 as treatment. VirtualPlant provides the user with several different statistical functions, but for this case study, select RankProduct (Breitling et al., 2004) with a $P$ value cutoff of 0.01. The calculation to determine differentially expressed genes is performed offline. An e-mail is sent to the user when the calculations are completed. It is not necessary to wait for this analysis to finish in order to do something else in VirtualPlant. For this case study, we also compared ATGE_81 versus ATGE_82, ATGE_82 versus ATGE_83, and ATGE_83 versus ATGE_84.

After completion of the statistical analysis, the user receives an e-mail notifying them that the job has been completed and indicating whether the analysis resulted in any differentially expressed genes. The list(s) of differentially expressed genes will appear in the cart with a name formed by concatenating the name of the experiment, "diff exp genes," the statistical method used, and whether the genes are induced (ind) or decreased (dec). Collectively, all four comparisons above will create eight lists (each comparison generating induced and repressed lists). A union of the eight lists will result in 1,367 genes that are differentially expressed in at least one of the four stages of seed development compared to the previous stage. To create a union of the lists, go to the "analysis view," select all the lists of differentially expressed genes during seed development you identified in the previous steps, and then select the "union" function from the analysis pull-down menu.

The next step in this case study is to create a molecular network for the genes that are regulated during seed development. Currently, VirtualPlant offers three different network functions: (1) super node networks, (2) gene networks, and (3) networks statistics. The super node networks analysis provides a view of the biological processes that are regulated

during seed development and how they interact with each other. The super node networks tool groups individual genes into a "super node" based on shared functional properties, such as GO terms, KEGG pathway, gene families, and even similar annotations. Edges are drawn between two super nodes when at least one gene or gene product in each super node has a molecular interaction. To perform super node networks analysis on the lists of differentially expressed genes during seed development, click on "analyze," select the check boxes next to the eight lists generated in the previous section (or the union of these lists), select the "super node networks" tool from the analysis pull-down menu, and click the "analyze" button. Once executed, the "super node networks" tool will present the user with two forms. The first form enables selection of the criteria for grouping genes into a super node (Fig. 5A). The default grouping method is to use the first few words of the gene annotations. For the first option, use the pull-down menu to select "share first TWO words," which will group together genes that share the first two words. The second option is to select the functional annotation you want to use. From the pull-down menu, select "KEGG pathway and gene families." Functional annotations are often categorized in a hierarchical manner, where the functional terms and pathways are themselves grouped into a higher more generic category. For the third option, select "direct associations" from the pull-down menu. Metabolic genes are often associated with each other via

A




B

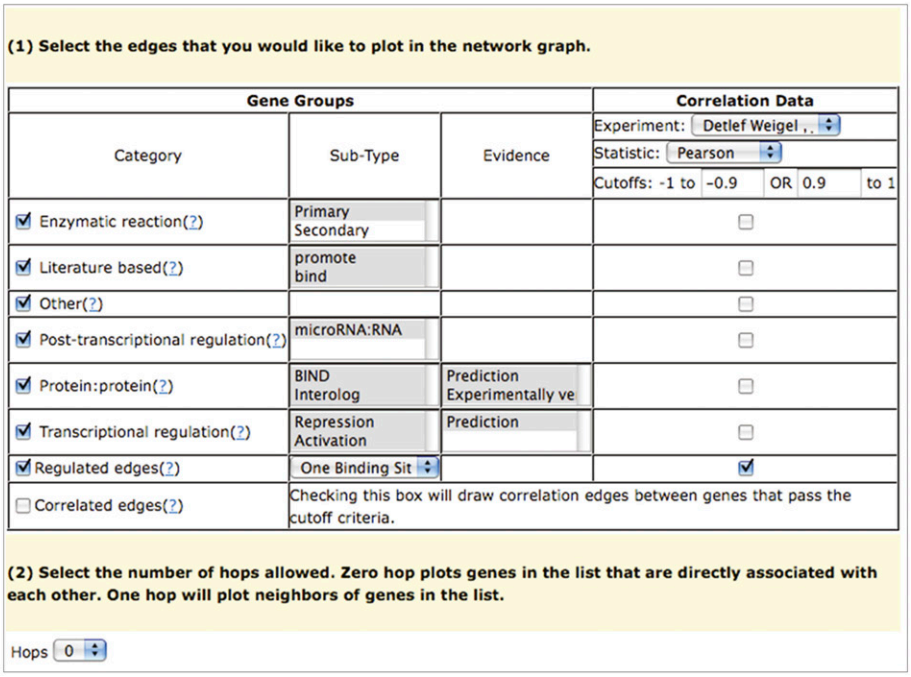



**Figure 5.** Super node and gene network forms. Super node analysis groups the genes based on the biological processes, functional terms, and annotations associated with the genes. A, The super node network form allows the user to choose from a selection of different functional term annotations and the depth of the annotation. In this case, the grouping is based on "KEGG pathway and gene families," and only the "direct associated" annotations are used. In the super node analysis, interactions between the biological processes are determined by the multinetwork data. Therefore, super node analysis will prompt the user with two forms: the super node network form and the multinetwork form. B, The gene network form allows the user to select from the different molecular interactions that are present in the multinetwork (see Table I for the list of resources available). In addition to the super node analysis, this form is also used for the network statistics tool.

metabolites. For this case study, we will not represent the metabolites in the network. Finally, click on the "submit" button. The next form allows the user to select the types of molecular interaction data to view in the network. See the "Materials and Methods" section for details about the different types of edges connecting two genes. Using the default mode will select all potential edges connecting two genes. For this case study, select enzymatic reactions, literature-based interactions, posttranscriptional regulation, protein-protein interactions, and transcriptional regulation (Fig. 5B). For enzymatic reactions, only select the "primary" reactions, which correspond to the edges drawn on the KEGG pathway maps. The regulated edges are predicted interactions based on the presence of known transcription factor cis-acting binding sites located in the 3-kb upstream region of annotated transcripts. Subtype "one binding site" represents presence of at least one binding site in the upstream region, and "over-represented binding site" represents overrepresentation of the binding site (two SDs) compared to the expected number based in all upstream regions in the genome. Check the "regulated edges" box and choose "one binding site" as the subtype. To improve the regulatory interaction predictions, filter the transcription factor:target gene predictions to include only the transcription factor and target pair whose expression values are correlated in the microarray experiment (Gutiérrez et al., 2008; Vandepoele et al., 2009). To filter "regulated edges" by correlation, select the checkbox in the "correlation data" column in the "regulated edges" row. To select the correct data set, select "Detlef Weigel, Jan Lohmann, Markus Schmid AtGenExpress: Developmental series (siliques and seeds)" in the "experiment" field. The statistics for the calculation of correlations selected in this example are "Pearson" and with cutoff values of less than −0.9 and higher than 0.9. The last parameter that we need to define to load the network is the number of "hops" away from the original list of genes used for the analysis. With 0 hops the network shows only the genes in the original list and the interactions between them. With 1 hop, the network will also show genes that were not in the original list but that are associated with genes in the original list. One hop is a good option when the gene list is small or has very few interactions. For this example, we will select 0 hops and then click on the "submit" button to generate the network. Visualization and manipulation of the network produced by either "super node analysis" or "gene network" analysis is implemented by the Cytoscape software (Shannon et al., 2003), which is launched automatically using Java Webstart. Features in Cytoscape allow users to set visualization preferences, such as the network layout (Figs. 6 and 7 use the organic layout), changing node attributes such as size (size of super nodes in Fig. 6 are proportional to the number of genes in the super node), and to select nodes based on attributes such as size. The first time Cytoscape is launched from VirtualPlant it will need to download the necessary files onto the user's computer. The super node networks analysis for the 1,367 seed regulated genes reveals several major transcription factor families that are highly connected in the seed

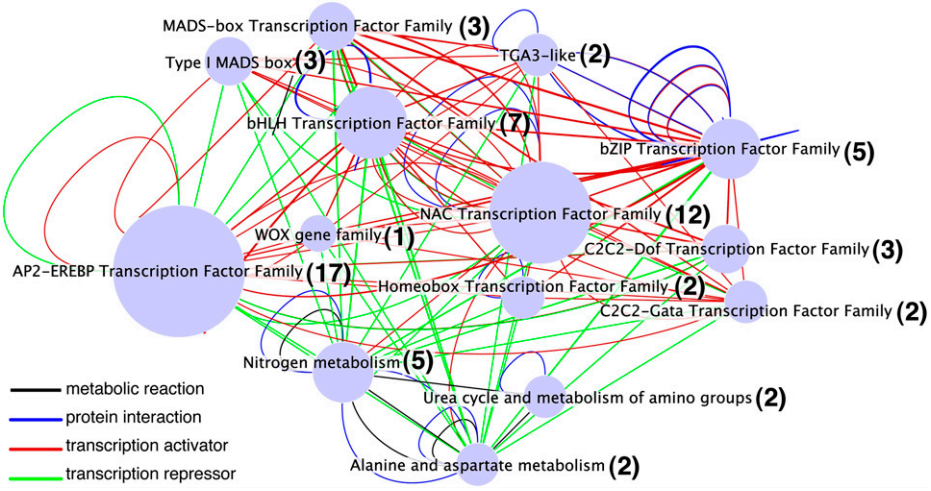



**Figure 6.** Super node network analysis of genes differentially expressed during seed development. The super node network graph allows the user to visualize relationships between biological processes. The nodes in the graph correspond to the super nodes, each grouping genes with common features, and edges connecting the nodes represent the different interactions between the genes in the super nodes (see text for details). Edge colors represent different interactions: blue edges, protein-protein interactions; black arrows, metabolic reactions; red arrows, predictions for transcriptional induction; and green arrows, predictions for transcriptional repression. The network shows "nitrogen metabolism" and its first neighbors in the super node seed-regulated network. The neighbors are mostly transcription factor families and two metabolic processes. The number near each name identifies the number of genes in the super node.

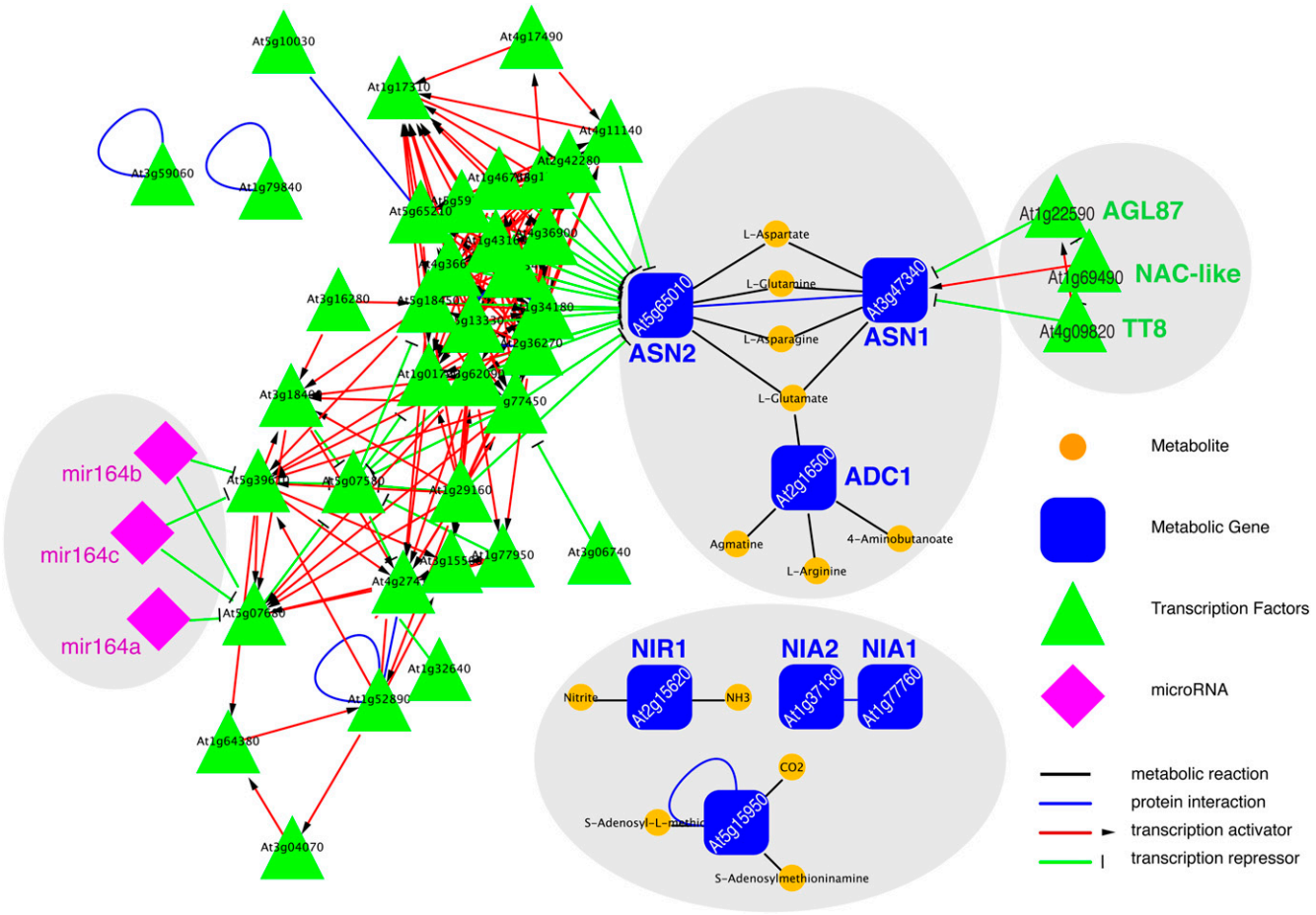



**Figure 7.** Gene network analysis of genes differentially expressed during seed development. The gene network graph shows interactions between genes, gene products, and/or metabolites. Orange circles represent metabolites, green triangles represent transcription factors, purple diamonds represent microRNAs, and blue squares represent metabolic genes. Edge colors represent different interactions: blue edges, protein-protein interactions; black arrows, metabolic reactions; red arrows, predictions for transcriptional induction; and green arrows, predictions for transcriptional repression. Different miR164 genes are shown targeting two transcription factors that are indirectly connected to the metabolic genes. Out of the seven nitrogen metabolic genes present in this network, only *ASN1* and *ASN2* have predicted regulators based on correlated transcription analysis and predicted cis-element binding sites.

development network (based on correlation and over-representation of cis-acting elements in the promoter region), including MADS box, bHLH, TGA3-like, NAC, and bZIP. Interestingly, the network analysis also identifies a super node of "unknown proteins" composed of 147 genes connected by putative transcription factor hubs. The next step is to identify the "nitrogen metabolism" node and all the other super nodes that are connected to it. From Cytoscape's menu bar, perform the following selection events: "select," "nodes," "by name," and type in the text field "nitrogen metabolism." This will highlight and select the node. From the Cytoscape's menu bar, select "select," "nodes," and "first neighbors of selected nodes." This will select the nitrogen metabolism node and all the nodes associated with it. Most of the neighbors are transcription factor families, and two are other metabolic process ("alanine and aspartate metabolism" and "urea cycle and metabolism of amino groups"; Fig. 6). The VirtualPlant plugin for Cytoscape allows users to send genes in the selected nodes back to their cart. While the nodes are still selected, from Cytoscape's menu bar select "plugins," "VirtualPlant," and then "login to VirtualPlant." Enter your VirtualPlant password and click the "login" button. Then again from the menu bar, select "plugins," "VirtualPlant," and "send selected nodes to VirtualPlant." A window will appear where the user must select the species they are working with. Select "Arabidopsis" and then "OK." Give the new list a name. There are 58 genes that are present in the super node nitrogen metabolism and the super nodes it interacts with, including transcription factors.

The Super node network analysis has identified a network of nitrogen metabolic genes and their neighbors that are regulated during seed development.

Nodes with the highest number of connections (hubs) in a biological network often play important roles in the network's operation (Barabasi and Oltvai, 2004). To obtain a quantitative measure of the number of connections, the user can run the "network statistics" tool. In the "analysis" view, select the checkbox near the list of 58 genes and select the "network statistics" function from the analysis pull-down menu. This tool displays a table of the most highly connected nodes in the network. The analysis revealed that several AP2-like transcription factors are among the most highly connected transcription factors in the nitrogen metabolic seed regulatory network, which suggests that they play an important role in regulating genes involved in nitrogen metabolism during the stages of seed development analyzed (Table II). To obtain a detailed view of all the molecular interactions of the 58 genes from the super node network analysis, e.g. individual transcription factors and their targets, the user can run the "gene networks" tool. From the analysis page, select the checkbox near the list of 58 genes and then "gene networks" from the analysis pull-down menu. Selecting the same options as before (primary enzymatic reactions, all the literature-based interactions, post-transcriptional regulation, protein-protein interactions, and transcriptional regulation) will produce the result shown in Figure 7. Five genes involved in nitrogen metabolism (*NIR1*, *NIA1*, *NIA2*, *ASN1*, and *ASN2*), three different miRNA164 genes, and 39 transcription factors from many different transcription factor families are present in the network. In this network, only two nitrogen metabolic genes are targeted by the transcription factors; *ASN1*, which is induced during seed development (Table III), and *ASN2*, which is repressed during seed development. The expression of *ASN1* is positively correlated to the expression of one transcription factor and negatively correlated to the expression of two other transcription factors. In contrast, the expression of *ASN2* is negatively correlated to the expression of all the transcription factors in the network.

**Table III.** *Five nitrogen metabolic genes that are regulated during seed development*

This table displays the five nitrogen metabolic genes and the stages of development in which they are regulated and how. IND, Induced; DEC, repressed.

| Genes | ATGE 79–81 | ATGE 81–82 | ATGE 82–83 | ATGE 83–84 |
|---|---|---|---|---|
| ASN1 | | IND | | |
| ASN2 | | DEC | | |
| NIA1 | | | IND | |
| NIA2 | | | IND | |
| NIR1 | IND | | | |

In our in silico network analysis, *ASN1* is one of the nitrogen metabolic genes that is regulated during seed development. Previous studies have shown that when *ASN1* is overexpressed using a 35S::*ASN1* line, the seed contains a higher level of free Asn (Lam et al., 2003). Along with higher levels of Asn, the authors also observed higher levels of total protein content in seeds. The results from our third case study also predict that the expression of *ASN1* is induced by a *NAC*-like transcription factor NAP (At1g69490), which itself is known to be required for leaf senescence (Guo

**Table II.** *Several AP2 transcription factors are highly connected in the seed development gene network*

This table ranks genes by the degree (number of connections) in the seed development gene network (see text for details).

| Gene | Connections | Annotation |
|---|---|---|
| At5g65010 | 23 | ASN2,ASN2 (ASPARAGINE SYNTHETASE2); Asn synthase (glutamine-hydrolyzing) |
| At2g36270 | 22 | ABI5,ABI5 (ABA INSENSITIVE5); DNA binding/ transcription factor/transcriptional activator |
| At5g18450 | 21 | AP2 domain-containing transcription factor, putative |
| At5g13330 | 20 | RAP2.6 L,RAP2.6 L (related to AP2 6L); DNA binding/ transcription factor |
| At1g34180 | 19 | ANAC016,ANAC016 (Arabidopsis NAC domain containing protein 16), ANAC016 (Arabidopsis NAC domain containing protein 16); transcription factor |
| At4g36900 | 18 | RAP2.10,RAP2.10 (related to AP2 10); DNA binding/ transcription factor |
| At3g62090 | 17 | PIL2,PIL2 (PHYTOCHROME INTERACTING FACTOR 3-LIKE2),PIL2 (PHYTOCHROME INTERACTING FACTOR 3-LIKE2); transcription factor |
| At1g01720 | 17 | ATAF1,ATAF1 (Arabidopsis NAC domain containing protein 2); transcription factor |
| At1g43160 | 17 | RAP2.6,RAP2.6 (related to AP2 6); DNA binding/ transcription factor |
| At1g77450 | 17 | ANAC032,ANAC032 (Arabidopsis NAC domain containing protein 32); transcription factor |

and Gan, 2006) and suggested to be involved in senescence of reproductive tissue (Kunieda et al., 2008). This result is consistent with the role of Asn in N-remobilization from leaves to developing seeds. The repressors of *ASN1* are *TT8* and *AGL87*. *TT8* was isolated while screening for seed coat color and is expected to play a role in flavonoid metabolism (Nesi et al., 2000). *AGL87* is a transcription factor from the MADS box family that has not been implicated during seed development. Little is known about *ASN2*, but this result supports a role for this gene during seed development. *ASN1* and *ASN2* are known to be reciprocally regulated, especially in light; thus, it is interesting to see *ASN1* and *ASN2* also regulated reciprocally during seed development (Table III). The network analysis hypothesizes that the five metabolic genes (*NIR1*, *NIA1*, *NIA2*, *ASN1*, and *ASN2*) are important in nitrogen metabolism during seed development, and it also proposes putative regulators of *ASN1* and *ASN2*, hypotheses that can be experimentally validated.

## DISCUSSION

Data interpretation, not data generation, has become an important bottleneck hindering the advancement of science. In an effort to help biologists take advantage of the burgeoning supply of genomic data, we have developed VirtualPlant, a Web site that enables scientists to integrate, analyze, and visualize genomic data to facilitate interpretation as well as generation of testable biological hypotheses. VirtualPlant implements and combines quantitative and visual approaches to data integration and analysis using a user-friendly, Web-accessible interface. The tools available from the VirtualPlant Web site (www.virtualplant.org) help biologists mine genomic data to address relevant questions in plant biology. Here, we have provided a series of case studies that demonstrate how a biologist can use VirtualPlant to analyze gene lists, gene networks, and microarray experiments. For a complete list of tools available in VirtualPlant, refer to the "help" section in the Web site.

An important feature of VirtualPlant is the cart for data storage and analysis. A user can store gene lists or experiments and execute tools that access the data stored in the cart. Most tools in VirtualPlant allow the user to save the results from the tool in the cart for further processing. The cart stores and organizes results indefinitely, so users can resume an analysis at any time. The iterative nature of analysis enabled by the cart helps filter and refine large data sets (gene lists, networks, or microarray experiments) to develop concrete testable biological hypotheses (Wang et al., 2004; Gutiérrez et al., 2007b, 2008; Gifford et al., 2008; Thum et al., 2008). Currently, the main data types the VirtualPlant system works on are lists of genes and microarray experiments. Our development efforts contemplate adding networks, metabolic pathways, and other complex data types to be handled directly by the VirtualPlant software.

When analyzing genomic data, biologists can often fail to discover interesting genes for experimental analysis when dealing with hundreds of putative candidate genes. They can also spend considerable time and effort copying and pasting lists of genes to perform such simple tasks as finding intersections between multiple lists. The first two case studies illustrate how a user can manage and analyze one or more lists of genes and easily perform set operations or the analysis of overrepresented functional terms. Our third case study demonstrates more advanced uses of VirtualPlant to analyze microarray data and to generate gene networks. In that example, we used VirtualPlant to identify gene networks and regulatory hubs that control seed development. A list of 1,367 genes that are regulated during seed development was obtained from the statistical analysis of publicly available microarray data. The combined use of two different network analysis tools led to a small set of AP2-like transcription factors that are predicted to act as regulatory hubs of nitrogen metabolism during seed development. The potentially key role of these transcription factors in nitrogen metabolism during seed development is supported by the phenotypes in the seeds of some of the mutant transcription factors and their targets. VirtualPlant allowed us to recapitulate existing knowledge about seed development, and it also allowed us to derive putative regulatory interactions, which may now be validated experimentally. Moreover, it allowed us to associate 147 genes of unknown function to seed development, thus prompting the hypothesis that these genes may have important functions during seed development.

With the widespread use of genomic technologies, the types of questions that are now common among biologists require a system that manages and analyzes sets of genes rather than individual genes. In addition, many biological processes are a result of interacting gene modules rather than isolated genes or gene products. The different types of data analysis (set operation, functional analysis, and gene networks) and visualization tools supported by VirtualPlant enable biologists to analyze genomic data from a systems perspective.

## MATERIALS AND METHODS

### VirtualPlant Software and Database Architecture

VirtualPlant is written in OO Perl using a model-view-controller design and other well-established patterns of software design. Data persistence in the VirtualPlant system is facilitated by the open source MySQL v5.0 database server. Our database schema uses a parsimonious design inspired by the LIMBO system (Philippi, 2004), with only four tables (OBJECT, OBJECT_CONNECTION, OBJECT_ATTRIBUTE, and CONNECTION_ATTRIBUTE) that support flexible accommodation of disparate data types. With careful attention to the indices and storage parameters of the database, we have found that this design provides high performance for a key set of queries that manages objects and their attributes and interconnections. Perl objects are transparently stored and retrieved from the database by a custom object relational mapping layer. A detailed description of the software and database

architecture is available upon request. VirtualPlant is freely available for use on the web and can be found at http://www.virtualplant.org. The source code is available upon request through a license agreement.

## VirtualPlant Data

Currently, the database contains most recently updated versions of (1) Arabidopsis (*Arabidopsis thaliana*) annotation from TAIR (ftp.arabidopsis.org; Rhee et al., 2003); (2) GO terms and their association to Arabidopsis genes (http://www.geneontology.org; Ashburner et al., 2000); (3) MipsFuncat functional categories and their association to Arabidopsis genes (ftp.mips.gsf.de; Mewes et al., 2004); (4) Affymetrix probes from ATH1 chips and their association to Arabidopsis genes downloaded from TAIR (ftp.arabidopsis.org); (5) our Multinetwork that is queried to create the network interaction discussed in case study 3, comprising biochemical pathways, including enzymes, reactions, and small molecules from KEGG (ftp.genome.jp; Kanehisa et al., 2004) and AraCyc (ftp.arabidopsis.org; Mueller et al., 2003); (6) protein interaction data from Bind (ftp.blueprint.org; Bader et al., 2004) and AtPID (Cui et al., 2008) databases, and experimentally determined protein interactions from Calmodulin (Popescu et al., 2007) and MADS BOX (de Folter et al., 2005) data sets; and (7) regulatory interaction data from the AGRIS database (Arabidopsis.med.ohio-state.edu/; Davuluri et al., 2003).

The VirtualPlant database also contains publicly available microarray data obtained from the NASC Affy Watch subscription (Craigon et al., 2004). The AtGenExpress (Schmid et al., 2005) and other widely used Arabidopsis microarray data sets are included in the NASC database of >1,800 hybridizations, performed using the Affymetrix AG and ATH1 DNA Chips. All hybridizations were normalized using RMA (Irizarry et al., 2003), provided by the BioConductor project (Gentleman et al., 2004). These normalized experiments are loaded into VirtualPlant to enable users to make comparisons across treatments. The normalized gene expression patterns across the approximately 1,800 chips in NASC were then correlated using the Spearman method (Samuels and Witmer, 2003), and the significant correlations ($P \leq 0.01$) were recorded and stored in the VirtualPlant database.

It is important to note that this network model does not currently contain all genes in the Arabidopsis genome. At present, the Arabidopsis network model contains 16,562 nodes, of which 13,960 are genes and 97,423 interactions described by Gutiérrez et al. (2007b). The number differences in the current version of VirtualPlant compared to the original 2007 publication is due to database updates, addition of new data sets, and refinement of the protein-protein interaction predictions. The different types of interactions present in this network are summarized in Table I. In this version of the database, protein-protein interactions are obtained from the Interactome project (Geisler-Lee et al., 2007) and the BIND database (Bader et al., 2002). Genes or gene products that cannot be associated to another gene in the genome by any known or predicted molecular interaction are not included in the model. "Regulated edge" predictions have been described previously (Gutiérrez et al., 2008). Briefly, consensus cis-acting motif sequences from AGRIS (Davuluri et al., 2003) were searched within the 3-kb upstream regions of all genes in the Arabidopsis genome using the DNA pattern search tool available on the RSA tools server (van Helden, 2003). Upstream regions were not allowed to overlap with coding region of the upstream gene. The motifs were also not allowed to overlap. Our predicted regulatory network contains 21,698,658 regulatory edges, where 1,187 transcription factors contain at least one binding site in the promoter region of 25,429 target genes. Surely not all of the predicted regulatory edges are valid. As discussed above, there are two methods of reducing the putative edges: (1) look for binding sites that are overrepresented compared to the genome, and (2) only consider regulatory edges that are also correlated across a given microarray experiment. Using the combination of both of the methods has proven useful in previous studies (Gutiérrez et al., 2008).

## Supplemental Data

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

**Supplemental Table S1.** Case Study 2: BioMaps results.

## ACKNOWLEDGMENTS

We thank Dr. Pamela J. Green and Dr. Blake Myers for microRNA interaction information. We thank other researchers who have contributed to the VirtualPlant software code: Chris Poultney, Ranjita Shankar Iyer, Varuni Prabhakar, Teresa Colombo, Jason Reisman, and Juan Manuel Cabello. We thank all the beta testers and especially Dr. Miriam Gifford, Dr. Karen Thum, Dr. Mariana Obertello, and Dr. Gabriel Krouk for helpful comments.

Received September 5, 2009; accepted November 29, 2009; published December 9, 2009.

## LITERATURE CITED

**Al-Shahrour F, Diaz-Uriarte R, Dopazo J** (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics **20:** 578–580

**Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, 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

**Bader G, Betel D, Hogue C** (2002) BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res **31:** 248–250

**Bader JS, Chaudhuri A, Rothberg JM, Chant J** (2004) Gaining confidence in high-throughput protein interaction networks. Nat Biotechnol **22:** 78–85

**Barabasi AL, Oltvai ZN** (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet **5:** 101–113

**Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G** (2004) GO:TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics **20:** 3710–3715

**Brady SM, Provart NJ** (2009) Web-queryable large-scale data sets for hypothesis generation in plant biology. Plant Cell **21:** 1034–1051

**Breitkreutz BJ, Stark C, Tyers M** (2003) Osprey: a network visualization system. Genome Biol **4:** r22.21–r22.24

**Breitling R, Armengaud P, Amtmann A, Herzyk P** (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett **573:** 83–92

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

**Crawford NM** (1995) Nitrate: nutrient and signal for plant growth. Plant Cell **7:** 859–868

**Cui J, Li P, Li G, Xu F, Zhao C, Li Y, Yang Z, Wang G, Yu Q, Li Y, et al** (2008) AtPID: *Arabidopsis thaliana* protein interactome database—an integrative platform for plant systems biology. Nucleic Acids Res **36:** D999–D1008

**Davuluri R, Sun H, Palaniswamy S, Matthews N, Molina C, Kurtz M, Grotewold E** (2003) AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC Bioinformatics **4:** 25

**de Folter S, Immink RG, Kieffer M, Parenicova L, Henz SR, Weigel D, Busscher M, Kooiker M, Colombo L, Kater MM, et al** (2005) Comprehensive interaction map of the *Arabidopsis* MADS box transcription factors. Plant Cell **17:** 1424–1433

**DeRisi JL, Iyer VR, Brown PO** (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science **278:** 680–686

**Eisen M, Spellman P, Brown P, Botstein D** (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA **95:** 14863–14868

**Endy D, Brent R** (2001) Modelling cellular behaviour. Nature **409:** 391–395

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

**Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al** (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol **5:** R80

**Gifford ML, Dean A, Gutiérrez RA, Coruzzi GM, Birnbaum KD** (2008) Cell-specific nitrogen responses mediate developmental plasticity. Proc Natl Acad Sci USA **105:** 803–808

**Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ** (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res **34:** D140–D144

**Guo Y, Gan S** (2006) AtNAP, a NAC family transcription factor, has an important role in leaf senescence. Plant J **46:** 601–612

**Gustafson AM, Allen E, Givan S, Smith D, Carrington JC, Kasschau KD** (2005) ASRP: the Arabidopsis Small RNA Project Database. Nucleic Acids Res **33:** D637–D640

**Gutiérrez RA, Gifford ML, Poultney C, Wang R, Shasha DE, Coruzzi GM, Crawford NM** (2007a) Insights into the genomic nitrate response using genetics and the Sungear Software System. J Exp Bot **58:** 2359–2367

**Gutiérrez RA, Lejay LV, Dean A, Chiaromonte F, Shasha DE, Coruzzi GM** (2007b) Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis. Genome Biol **8:** R7

**Gutiérrez RA, Shasha DE, Coruzzi GM** (2005) Systems biology for the virtual plant. Plant Physiol **138:** 550–554

**Gutiérrez RA, Stokes TL, Thum K, Xu X, Obertello M, Katari MS, Tanurdzic M, Dean A, Nero DC, McClung CR, et al** (2008) Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene CCA1. Proc Natl Acad Sci USA **105:** 4939–4944

**Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J** (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics **22:** 2825–2827

**Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L** (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science **292:** 929–934

**Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP** (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4:** 249–264

**Kahlem P, Birney E** (2007) ENFIN a network to enhance integrative systems biology. Ann N Y Acad Sci **1115:** 23–31

**Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M** (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res **32:** D277–D280

**Kao HL, Gunsalus KC** (2008) Browsing multidimensional molecular networks with the generic network browser (N-Browse). Curr Protoc Bioinformatics **Chapter 9:** 11

**Karp PD** (1996) A strategy for database interoperation. J Comput Biol **2:** 573–586

**Khatri P, Draghici S, Ostermeier GC, Krawetz SA** (2002) Profiling gene expression using Onto-Express. Genomics **79:** 266–270

**Kunieda T, Mitsuda N, Ohme-Takagi M, Takeda S, Aida M, Tasaka M, Kondo M, Nishimura M, Hara-Nishimura I** (2008) NAC family proteins NARS1/NAC2 and NARS2/NAM in the outer integument regulate embryogenesis in *Arabidopsis*. Plant Cell **20:** 2631–2642

**Lam HM, Wong P, Chan HK, Yam KM, Chen L, Chow CM, Coruzzi GM** (2003) Overexpression of the ASN1 gene enhances nitrogen status in seeds of Arabidopsis. Plant Physiol **132:** 926–935

**Loew LM, Schaff JC** (2001) The Virtual Cell: a software environment for computational cell biology. Trends Biotechnol **19:** 401–406

**Lu C, Tej SS, Luo S, Haudenschild CD, Meyers BC, Green PJ** (2005) Elucidation of the small RNA component of the transcriptome. Science **309:** 1525–1526

**Mendes P** (1997) Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. Trends Biochem Sci **22:** 361–363

**Mewes HW, Amid C, Arnold R, Frishman D, Güldener U, Mannhaupt G, Münsterkötter M, Pagel P, Strack N, Stümpflen V, et al** (2004) MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res **32:** D41–D44

**Mueller LA, Zhang P, Rhee SY** (2003) AraCyc: a biochemical pathway database for Arabidopsis. Plant Physiol **132:** 453–460

**Nesi N, Debeaujon I, Jond C, Pelletier G, Caboche M, Lepiniec L** (2000) The TT8 gene encodes a basic helix-loop-helix domain protein required for expression of DFR and BAN genes in *Arabidopsis* siliques. Plant Cell **12:** 1863–1878

**Philippi S** (2004) Light-weight integration of molecular biological databases. Bioinformatics **20:** 51–57

**Popescu SC, Popescu GV, Bachan S, Zhang Z, Seay M, Gerstein M, Snyder M, Dinesh-Kumar SP** (2007) Differential binding of calmodulin-related proteins to their targets revealed through high-density Arabidopsis protein microarrays. Proc Natl Acad Sci USA **104:** 4730–4735

**Poultney CS, Gutiérrez RA, Katari MS, Gifford ML, Paley WB, Coruzzi GM, Shasha DE** (2007) Sungear: interactive visualization and functional analysis of genomic datasets. Bioinformatics **23:** 259–261

**Redman J, Haas B, Tanimoto G, Town C** (2004) Development and evaluation of an Arabidopsis whole genome Affymetrix probe array. Plant J **38:** 545–561

**Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, 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

**Ritter O** (1994) The integrated genomic database (IGD). *In* S Suhai, ed, Computational Methods in Genome Research. Plenum Press, New York, pp 57–73

**Rzhetsky A, Iossifov I, Koike T, Krauthammer M, Kra P, Morris M, Yu H, Duboué PA, Weng W, Wilbur WJ, et al** (2004) GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. J Biomed Inform **37:** 43–53

**Samuels ML, Witmer JA** (2003) Statistics for Life Science. Pearson Education, San Francisco

**Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Schölkopf B, Weigel D, Lohmann JU** (2005) A gene expression map of *Arabidopsis thaliana* development. Nat Genet **37:** 501–506

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

**Siepel A, Farmer A, Tolopko A, Zhuang M, Mendes P, Beavis W, Sobral B** (2001) ISYS: a decentralized, component-based approach to the integration of heterogeneous bioinformatics resources. Bioinformatics **17:** 83–94

**Sitt M, Muller C, Matt P, Gibon Y, Carillo P, Morcuende R, Scheible WR, Krapp A** (2002) Steps towards an integrated view of nitrogen metabolism. J Exp Bot **53:** 959–970

**Thum KE, Shin MJ, Gutiérrez RA, Mukherjee I, Katari MS, Nero D, Shasha D, Coruzzi GM** (2008) An integrated genetic, genomic and systems approach defines gene networks regulated by the interaction of light and carbon signaling pathways in Arabidopsis. BMC Syst Biol **2:** 31

**Usadel B, Obayashi T, Mutwil M, Giorgi FM, Bassel GW, Tanimoto M, Chow A, Steinhauser D, Persson S, Provart NJ** (2009) Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. Plant Cell Environ **32:** 1633–1651

**Vandepoele K, Quimbaya M, Casneuf T, De Veylder L, Van de Peer Y** (2009) Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks. Plant Physiol **150:** 535–546

**van Helden J** (2003) Regulatory sequence analysis tools. Nucleic Acids Res **31:** 3593–3596

**Wang R, Tischner R, Gutiérrez RA, Hoffman M, Xing X, Chen M, Coruzzi G, Crawford NM** (2004) Genomic analysis of the nitrate response using a nitrate reductase-null mutant of Arabidopsis. Plant Physiol **136:** 2512–2522

**Wilkinson M, Schoof H, Ernst R, Haase D** (2005) BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. The PlaNet exemplar case. Plant Physiol **138:** 5–17

**Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, et al** (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol **4:** R28

**Zhong S, Tian L, Li C, Storch FK, Wong WH** (2004) Comparative analysis of gene sets in the Gene Ontology space under the multiple hypothesis testing framework. Proc IEEE Comput Syst Bioinform Conf **2004:** 425–435