This article is a *Plant Cell* Advance Online Publication. The date of its first appearance online is the official date of publication. The article has been edited and the authors have corrected proofs, but minor changes could be made before the final version is published. Posting this version online reduces the time to publication by several weeks.

REVIEW

Web-Queryable Large-Scale Data Sets for Hypothesis Generation in Plant Biology

Siobhan M. Brady[a] and Nicholas J. Provart[b,1]

[a] Section of Plant Biology and Genome Center, University of California, Davis, California 95616
[b] Department of Cell and Systems Biology/Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, ON M5S 3B2, Canada

The approaching end of the 21st century's first decade marks an exciting time for plant biology. Several National Science Foundation *Arabidopsis* 2010 Projects will conclude, and whether or not the stated goal of the National Science Foundation 2010 Program—to determine the function of 25,000 *Arabidopsis* genes by 2010—is reached, these projects and others in a similar vein, such as those performed by the AtGenExpress Consortium and various plant genome sequencing initiatives, have generated important and unprecedented large-scale data sets. While providing significant biological insights for the individual laboratories that generated them, these data sets, in conjunction with the appropriate tools, are also permitting plant biologists worldwide to gain new insights into their own biological systems of interest, often at a mouse click through a Web browser. This review provides an overview of several such genomic, epigenomic, transcriptomic, proteomic, and metabolomic data sets and describes Web-based tools for querying them in the context of hypothesis generation for plant biology. We provide five biological examples of how such tools and data sets have been used to provide biological insight.

INTRODUCTION

The study of plant biology, as with all areas of biology, has undergone dramatic changes in the past decade. The development of technologically advanced, high-throughput methods for querying the expression levels of thousands of genes at once, for detecting interactions between proteins in a plant's proteome, or for simultaneously measuring the amounts of many metabolites has permitted unprecedented insight into many aspects of plant biology. Thousands of data sets encompassing millions of measurements have been generated, and importantly, most of these are freely available for use by any plant biologist worldwide to examine in the context of his or her biological question. While such large scale data sets may not provide complete understanding of a particular question, they are often an excellent starting point for planning experiments or generating hypotheses in silico or helping to make sense of one's own high-throughput data sets. These hypotheses can then be readily tested in the laboratory with the amazing variety of genetic resources and molecular techniques that have also been developed in the past 10 years.

This review provides an overview of the breadth and depth of data sets that are currently available, especially for, but not limited to, the model plant *Arabidopsis thaliana*. Many of these data sets were generated by researchers funded through the National Science Foundation *Arabidopsis* 2010 project in the U.S., the stated goal of which was to identify the functions of 25,000 genes in *Arabidopsis* by 2010 (Chory et al., 2000), and by the AtGenExpress Consortium, an international effort to uncover the *Arabidopsis* transcriptome. In this review, we emphasize Web-based tools that have integrated data from several sources. While many individual researchers have set up websites for their own data sets, resources that compare diverse data sets are often of more utility to a wider biological research audience. We describe well-developed sequence databases, focusing on transcriptome data sets, which are the most comprehensive of all of the large-scale data types, and discuss tools for querying these both in a directed manner and correlatively, using data mining tools for generating hypotheses or narrowing down search space. We also discuss databases of epigenetic modifications and small RNAs and survey metabolomic and proteomic resources. Tools for integrating disparate data types to improve function prediction are key to leveraging even more knowledge from these data sets, and two such tools will be reviewed. We conclude with some perspectives on what the future will bring in terms of queryable browsers for further understanding the plant as a collection of cellular systems and processes and of plant varieties at an ecophysiological level. Throughout this review, we provide bioexamples of how such large scale data sets have been used to expand our understanding of the processes described above, often at the cost of only a click of the mouse. An overview of the use of these tools and data sets for plant biology is given in Figure 1, and programs and websites discussed in this review are listed in Table 1.

Sequence Databases I: Genome Browsers

Gramene

Once a gene of interest has been identified, several logical questions arise, such as whether an ortholog exists for it in another

[1] Address correspondence to provart@utoronto.ca.
www.plantcell.org/cgi/doi/10.1105/tpc.109.066050.

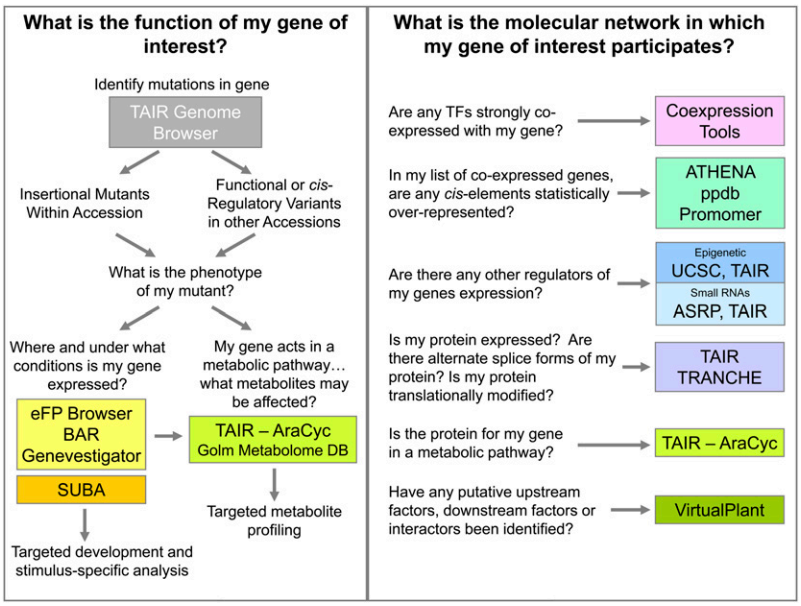

Figure 1. How Can Queryable Browsers Be Used to Address Biological Questions?

Queryable browsers are represented in colored boxes. Left panel: How queryable browsers can be used to elucidate the function of a gene of interest. Right panel: How queryable browsers can be used to elucidate the molecular network within which a gene of interest participates.

plant species, if the gene neighborhood is conserved in other species, or if there are polymorphisms that affect the coding region in other accessions. The user-friendly, Web-based Gramene Genome Browser (www.gramene.org) was developed as a resource for comparative genomics of grass species. Gramene uses the sequenced rice (*Oryza sativa*) genome as a scaffold to order and orient partially sequenced genomes based on their synteny to rice and as a reference to discover candidate genes in other crops (Liang et al., 2008). Full genome sequences from *O. sativa* ssp *japonica* cv Nipponbare, *A. thaliana*, and poplar (*Populus trichocarpa*) are accessible, as well as sequences from additional rice species (*O. sativa* ssp indica, *Oryza rufipogon*, and *Oryza glaberrima*), *Arabidopsis lyrata*, the grasses *Zea mays* and *Sorghum bicolor*, and the common grapevine *Vitis vinifera*. Gene tracks in the Gramene Genome Browser include gene structure visualized at its respective genomic location and neighboring loci. Tracks for non-protein-coding rice genes and protein-coding sequences annotated from a variety of species are also available. Syntenic genomic alignment can also be viewed for all available species' genomic sequences. Gene trees that show available putative orthologs and paralogs are also displayed and denoted by species and type (1-to-1, or 1-to-many). Tracks that display quantitative trait loci (QTL), single nucleotide polymorphisms (SNPs), ontology annotations, BLAST, and links to literature are also available (Liang et al., 2008). Gramene guides users via module tutorials, and Ware (2007) provides a working example of how a maize researcher can use Gramene for targeted experimental research.

The Arabidopsis Information Resource

The generation of the first genome sequence for the model plant *A. thaliana* in 2000 (Arabidopsis Genome Initiative, 2000) was a landmark for plant biology. Several iterations of annotation, currently at version 8 (TAIR8) with TAIR9 about to be released, have resulted in a mature and well-annotated genome sequence. Historically, research in *Arabidopsis* was confined to the use of a limited number of ecotypes, or inbred stocks, for which genetic maps and sequences were available. For many molecular studies, this was sufficient. In recent years, however, high-density oligonucleotide resequencing microarrays and next-generation sequencing technologies have resulted in a considerable increase in the amount of genome sequence data for this species. This has enabled evolutionary studies of adaptation and natural selection at the molecular level using genetically diverse natural accessions to show adaptation across a specific geographic range (Mitchell-Olds and Schmitt, 2006). The TAIR Genome Browser (www.Arabidopsis.org/cgi-bin/gbrowse/) has a host of tracks built into it. We will highlight several of these that allow the user to take advantage of currently available data as well as query genome sequences (Swarbreck et al., 2008).

The TAIR Genome Browser allows for visualization of multiple windows of sequence information within a chromosomal region. Useful windows include annotation units (genomic clones) that make up a tiling path, assembled chromosomes with chromosomal locations, the approximate position of available transposon insertion mutants, locus and protein gene coding models (including coding segments for each splice variant), cDNAs and EST sequences from GenBank, and plant gene family clusters (for poplar, *V. vinifera*, *O. sativa*, *Physcomitrella patens*, *S. bicolor*, *Selaginella moellendorffii*, and *Chlamydomonas reinhardtii*). Alignments of sequences to cDNAs and ESTs of *Brassica*, a closely related genus of *Arabidopsis*, are also available. Of particular note, positions of many different types of polymorphisms, some identified in different accessions, may also be

Table 1. Programs and URLs Discussed in This Review

Program	URL	Comments	Reference
Primarily sequence resources			
Gramene	www.gramene.org	Resource for comparative genomics in grass species; assembled genome sequence for grass species and for *A. thaliana*, *A. lyrata*, *V. vinifera*, and *P. trichocarpa*.	Liang et al. (2008)
TAIR Genome Browser	www.Arabidopsis.org/cgi-bin/gbrowse/	View polymorphisms, insertional mutant locations, cDNAs, ESTs, plant gene family clusters, and splice variants.	Swarbreck et al. (2008)
TIGR Gene Indices	compbio.dfci.harvard.edu/tgi/plant.html	TIGR Gene Indices (EST collections) for 45 plant species from apple to wheat are a rich resource for sequence information; searchable by BLAST.	Quackenbush et al. (2000)
SIGnAL	signal.salk.edu	The Salk Institute Genomic Analysis Laboratory website contains a wealth of information, including *Arabidopsis* genome sequence, transcriptome, epigenome, methylome, small RNA, and exosome substrate maps; functional genomic data for rice; provides a comprehensive listing of T-DNA insertions in *Arabidopsis*.	Alonso et al. (2003)
VISTA	genome.lbl.gov/vista/	The JGI's VISTA browser provides convenient cross-species comparison for the sequenced plant genomes of *Arabidopsis*, poplar, rice, *P. patens*, and *S. moellendorffii*.	Frazer et al. (2004)
***Cis*-element resources**			
ATHENA	www.bioinformatics2.wsu.edu/Athena/	Mapping of known *cis*-elements from several databases onto *Arabidopsis* promoters; enrichment analysis tools.	O'Connor et al. (2005)
AGRIS	Arabidopsis.med.ohio-state.edu	*Arabidopsis cis*-regulatory database and transcription factor database.	Davuluri et al. (2003)
PLACE	www.dna.affrc.go.jp/PLACE/	469 *cis*-elements, mainly from vascular plants, with cross-references to original articles describing them.	Higo et al. (1999)
PlantCARE	bioinformatics.psb.ugent.be/webtools/plantcare/html/	435 plant transcription sites: 149 from monocots, 281 from dicots, and 5 from other plants, describing >159 plant promoters.	Lescot et al. (2002)

(Continued)

Table 1. (continued).

Program	URL	Comments	Reference
AthaMap	www.athamap.de	Provides a genome-wide map of potential transcription factor and small RNA binding sites in *Arabidopsis*; searchable for combinatorial *cis*-element effects.	Steffens et al. (2005)
ppdb	ppdb.gene.nagoya-u.ac.jp	Annotated transcription start sites and regulatory elements of *Arabidopsis* and rice promoters. Also incorporated into TAIR's Gbrowse.	Yamamoto and Obokata (2008)
Gene expression resources			
BAR	BAR.utoronto.ca	Browse AtGenExpress and poplar expression data sets with the e-Northern tool or eFP Browser; perform coexpression studies and *cis*-element prediction; view precomputed CAPS markers with MarkerTracker and predicted interactions with AIV; view subcellular localization with Cell eFP Browser.	Toufighi et al. (2005), Geisler-Lee et al. (2007), Winter et al. (2007), Wilkins et al. (2008)
Genevestigator	www.genevestigator.com	Browse *Arabidopsis*, rice, barley, and soybean expression data; identify biomarkers; map to pathways, clustering tools.	Zimmermann et al. (2004, 2005), Grennan (2006), Hruz et al. (2008)
At-TAX	www.weigelworld.org/resources/microarray/at-tax/	Browse developmental and stress series expression data sets generated using whole-genome tiling arrays.	Laubinger et al. (2008)
Coexpression Tools	ATTEDII, Expression Angler, Genevestigator, AthCoR@CSB.DB, ACT, CressExpress, GeneCAT	These tools for identifying coexpressed genes are well described in the cited review by Aoki et al. (2007). CressExpress and GeneCAT are recent additions.	Aoki et al. (2007), Mutwil et al. (2008), Srinivasasainagendra et al. (2008)
Small RNA and epigenetic modification resources			
Arabidopsis Small RNA Project Database	asrp.cgrb.oregonstate.edu	Database for recent deep-sequencing projects cataloguing small RNAs in *Arabidopsis*.	Gustafson et al. (2005)
UCSC Genome Browser	epigenomics.mcdb.ucla.edu/H3K27m3/	View H3K27me3 methylation patterns generated by the Jacobsen/Pellegrini groups for *Arabidopsis.*	Zhang et al. (2007)
UCSC Genome Browser	epigenomics.mcdb.ucla.edu/BS-Seq/	Cytosine methylation patterns generated by the Jacobsen/Pellegrini groups at base pair resolution for *Arabidopsis*.	Cokus et al. (2008)
MPSS Database	http://mpss.udel.edu/at/	Explore RNA degradome data for *Arabidopsis* and MPSS expression data from several tissues for *Arabidopsis* (and rice).	Nakano et al. (2006), German et al. (2008)

(Continued)

Table 1. (continued).

Program	URL	Comments	Reference
Proteome resources			
NASC *Arabidopsis* Proteomics Database	proteomics.Arabidopsis.info	Two chloroplast proteomics experiments, one using DIGE on wild-type and mutant chloroplasts, and one using LOPIT, may be queried.	Kubis et al. (2003), Dunkley et al. (2006)
GABI PD	www.gabipd.org/projects/*Arabidopsis*_Proteomics/	A handful of developmental stages in *Arabidopsis* and *Brassica rapus* as examined by 2D gel electrophoresis may be explored.	Riano-Pachon et al. (2009)
PRIDE, AtProteome	www.ebi.ac.uk/pride/prideMart.do, fgcz-atproteome.unizh.ch	Query a proteomic database examining the presence of proteins in six *Arabidopsis* organs.	Baerenfaller et al. (2008)
PhosPhAt	phosphat.mpimp-golm.mpg.de	6282 phosphopeptides (5948 of these are from 10 publications). Queryable by AGI ID or by peptide.	Heazlewood et al. (2008)
SUBA	plantenergy.uwa.edu.au/suba2/	Documented localizations of >6000 *Arabidopsis* proteins and predicted localizations for most.	Heazlewood et al. (2007)
Cell eFP Browser	BAR.utoronto.ca	Pictographic display of SUBA subcellular localization data.	Winter et al. (2007)
Arabidopsis Interactions Viewer	BAR.utoronto.ca	Display of ~70,000 predicted protein–protein interaction data by Geisler-Lee et al. (2007) and ~2800 documented *Arabidopsis* protein–protein interactions.	Geisler-Lee et al. (2007)
AtPID	atpid.biosino.org	Queryable database of ~28,000 documented and predicted interactions in *Arabidopsis*.	Cui et al. (2008)
Metabolome resources			
BinBase	http://eros.fiehnlab.ucdavis.edu:8080/binbase-compound/	Documentation of >1000 well-characterized small molecules in several plant species.	Fiehn et al. (2005)
Golm Metabolome DB	csbdb.mpimp-golm.mpg.de/csbdb/gmd/profile/gmd_smpq.html	Search for several hundred identified metabolites from plants grown under different light conditions.	Kopka et al. (2005)
Integrative resources			
VirtualPlant	www.VirtualPlant.org	Integrate several disparate types of data from *Arabidopsis* for identifying novel components of a given system.	Coruzzi et al. (2006)
GeneMANIA	morrislab.med.utoronto.ca/mania/	Integrate several disparate types of data from *Arabidopsis* for identifying novel components of a given system.	Mostafavi et al. (2008)

visualized. Such polymorphisms can be used in functional genetic analyses, in the identification of causal alleles found from QTL studies, and in studies of evolutionary processes that shape population-wide sequence variation. Polymorphisms include TILLing mutations, SNPs, and insertions and deletions (Clark et al., 2007; Ossowski et al., 2008; Zeller et al., 2008).

SIGnAL, TIGR Gene Indices, and VISTA

For other plant species, TIGR Gene Indices (Quackenbush et al., 2000) represent a rich collection of assembled ESTs (these are called tentative consensus sequences, or TCs) across 45 species from apple (*Malus domestica*) through sugarcane (*Saccharum officinarum*) to wheat (*Triticum aestivum*). Several species' tentative consensus sequences are displayed in the TAIR Genome Browser. The Gene Indices, now housed at the Dana Farber Cancer Institute at Harvard, are also searchable by BLAST directly at compbio.dfci.harvard.edu/tgi/plant.html. The Salk Institute's SIGnAL website (signal.salk.edu) also provides a genome browser to view *Arabidopsis* and rice T-DNA insertion lines, expression data, and orthologs, along with several tools for querying these data (Alonso et al., 2003). Finally, the Department of Energy's Joint Genome Institute offers the VISTA Genome Browser (genome.lbl.gov/vista/) for exploring five sequenced plant genomes: *Arabidopsis*, rice, poplar, *P. patens*, and *S. moellendorffii* (Frazer et al., 2004).

Sequence Databases II: *cis*- and Regulatory Element Databases and Browsers

For a given gene of interest, or for a set of genes that share similar expression patterns, a common question is whether the promoter or promoters contain known *cis*-acting elements that are responsible for directing gene expression in a particular manner. *Cis*-element databases and tools for exploring these can be considered a subset of sequence databases. ATHENA (O'Connor et al., 2005), AGRIS (Davuluri et al., 2003), PLACE (Higo et al., 1999), PlantCARE (Lescot et al., 2002), and AthaMap (Steffens et al., 2005) are the major repositories for plant *cis*-regulatory elements (see Table 1 for URLs). Unfortunately, the updating of at least one of these, PLACE, has been discontinued since February 2007. AGRIS does not appear to have been updated since 2004, although a forthcoming update promises to double the number of documented *cis*-elements in the database. Recently, TAIR's Genome Browser started incorporating data from ppdb, ppdb.gene.nagoya-u.ac.jp, a plant promoter database that annotates *Arabidopsis* and rice promoter structure, including both novel and already characterized transcription start sites and regulatory elements (Yamamoto and Obokata, 2008). Regulatory sequences are linked to the literature and to other promoters containing the same sequence (Yamamoto and Obokata, 2008). Studies aimed at generating sequence from all *Arabidopsis* full-length cDNAs have refined predicted transcriptional start sites (Seki et al., 2002). Iida et al. (2004) have used this information to identify genome-wide alternative pre-mRNA splicing events in this species. Although databases with a limited number of *cis*-regulatory sequences are currently available, a comprehensive listing of *cis*-regulatory elements and their cognate transcription

Bioexample 1: Deep Sequencing to Explore Polymorphisms That Shape Natural Variation in *Arabidopsis*

High-density oligonucleotide resequencing microarrays have been used to determine the types of polymorphisms that exist among 20 accessions with maximal genetic diversity (Clark et al., 2007; Zeller et al., 2008). Using this technology and machine learning methods, short polymorphic tracts of <10 bp in size and extended polymorphic tracts, including long deletions, were identified and have been included in the TAIR Genome Browser (Zeller et al., 2008). Nearly 10% of all protein-coding genes were identified to contain large-effect SNPs (premature stop codons, altered initiation Met residues, and nonfunctional splice donor or acceptor sites), demonstrating significant sources of potential functional variation across these accessions (Clark et al., 2007). Patterns of sequence variation were also assessed for gene families. Nucleotide binding leucine-rich repeat genes that mediate disease resistance and F-box genes that act in ubiquitin-mediated protein degradation show extreme levels of polymorphism, while transcription factors and microRNA (miRNA) loci show little variation (Clark et al., 2007; Zeller et al., 2008). Allele frequency patterns in the SNP data suggest balancing selection as an evolutionary force leading to high polymorphism levels for the nucleotide binding leucine-rich repeat family (Clark et al., 2007; Zeller et al., 2008). When the polymorphism data were used to infer the distribution of polymorphisms in intergenic sequences, polymorphisms varied as a function of distance from coding sequences; the number of polymorphisms falls drastically starting ~450 bp upstream of the start of the coding region, within the 5′ untranslated region (Zeller et al., 2008). This drop in polymorphic sequence as identified by resequencing arrays suggests that much deeper sequencing is required to identify functional *cis*-regulatory variants that might play a functional role in environmental adaptation in closely related *Arabidopsis* species (Hanikenne et al., 2008).

Discoveries revealed by the array-based resequencing approach and the emergence of low-cost, high-throughput sequencing technology have motivated the 1001 Genomes project (1001genomes.org), whose goal is to sequence the genomes of 1001 *Arabidopsis* accessions. Ultimately, sequencing of this many genomes will greatly facilitate genome-wide association mapping by increasing our ability to map causal variants responsible for QTL at the nucleotide level. As a proof of principle, two divergent accessions, Bur-0 and Tsu-1, have been sequenced using this method. A total of 823,325 unique SNPs and 79,961 unique 1- to 3-bp indels (insertion or deletion mutations) were identified, with 15- to 25-fold coverage in reads (Ossowski et al., 2008). These polymorphisms have been incorporated into the TAIR Genome Browser. The methods for aligning reads and for predicting SNPs and indels will be used for further accession sequencing (Ossowski et al., 2008). Identification of further major effect changes in protein-coding sequences will greatly facilitate future functional studies within these diverse accessions. Keep a close eye on the 1001 Genomes website in the future!

factors is lacking, primarily due to a deficiency in experimental validation. Perhaps what is needed is a systematic project to determine the binding specificities of all transcription factors in *Arabidopsis* in a manner similar to that which has been performed for 168 mouse transcription factor homeodomains using universal DNA microarrays encompassing all possible 8-mers (Berger et al., 2006, 2008).

Gene Expression Databases and Browsers

Coming back to one's gene or genes of interest, a suitable next question is, what is the expression pattern of my gene of interest? Expression patterns can then be used to guide further biological experiments. Additional questions might include: within my gene's family, are family members uniquely expressed in certain tissues or is one uniquely upregulated by a specific abiotic or biotic stress suggesting subfunctionalization or neofunctionalization? Are there other genes, not currently known to be involved in my gene's given biological process, that exhibit similar patterns of expression? Alternately, if one is not aware of a given gene's biological function, what are the functions of other genes that are similarly expressed with the gene of interest? Moving away from a single gene-centered approach, one might also ask what is the full set of transcriptional programs that occur in my tissue of interest or response condition?

Originally, genome-wide expression measurements were limited to the use of cDNA or EST resources. Seki et al. (2004) used microarrays containing RIKEN *Arabidopsis* full-length cDNA sequences to identify many novel abiotic stress-induced genes. Of course, without the availability of a genome and cDNA sequences, the development of the widely used short oligonucleotide microarrays for measuring the transcriptome of *Arabidopsis* would not have been possible. Affymetrix's 8K At GeneChip (Zhu and Wang, 2000), subsequent 22K ATH1 GeneChip (Redman et al., 2004), and *Arabidopsis* Tiling 1.0R Array (Laubinger et al., 2008) all have been used to examine the transcriptomes of bulk tissues and specific cell types, both within the framework of the international AtGenExpress project and by individual researchers. Data sets associated with the AtGenExpress project profile a wide variety of developmental stages, tissues, cell types, hormone responses, and biotic and abiotic stresses (Schmid et al., 2005; Kilian et al., 2007; Goda et al., 2008). These extensive resources can be mined to generate hypotheses. Mining of such data can also result in the elucidation of putative transcriptional modules by identifying genes coexpressed with a gene of interest, in the inference of cell type-, tissue-, or context-specific expression of genes within large, seemingly redundant families and in the identification of genes potentially acting within complexes. More than 4400 data sets generated with the Affymetrix ATH1 platform have been deposited to GEO (the National Center for Biotechnology Information [NCBI] Gene Expression Omnibus) at www.ncbi.nlm.nih.gov/geo/ (Edgar et al., 2002), and these may be downloaded for further analysis by independent researchers using the open source BioConductor suite (Gentleman et al., 2004). Two Web-queryable databases, which have incorporated more than half of the these expression data sets and provide many useful tools, are commonly used within the *Arabidopsis* community due to their user-friendly interfaces and data mining capabilities: the Bio-Array Resource for *Arabidopsis* Functional Genomics (the BAR; BAR.utoronto.ca) and Genevestigator (www.genevestigator.com) (Zimmermann et al., 2004, 2005; Toufighi et al., 2005; Grennan, 2006; Geisler-Lee et al., 2007; Winter et al., 2007; Hruz et al., 2008; Wilkins et al., 2008). The BAR and Genevestigator provide many tools for analysis of *Arabidopsis* microarray expression data and for expression data from other species. The BAR additionally allows investigation of mouse, poplar, and *Medicago truncatula* expression data, while Genevestigator allows investigation of human, mouse, rat, barley (*Hordeum vulgare*), rice, and soybean (*Glycine max*). We will briefly describe the tools available on both these websites, highlighting unique aspects of each.

The BAR

Analysis tools at the BAR include the Expression Angler, Expression Browser, the electronic Fluorescent Pictograph (eFP) Browser, and Promomer tools. The Expression Browser tool takes as its input large lists of genes and queries expression across user-selected expression data sets, thus allowing gene expression levels to be determined during development or in response to stresses. These data can be displayed in plain text or hierarchically clustered and visualized (Toufighi et al., 2005). Of particular utility is the user's ability to output either absolute expression levels in various treatment and control samples or the ratio of response level in the treatment relative to the level in the control. Gene expression can be visualized using the eFP Browser tool. Here, expression of one or two genes is visualized in stylized pictographs of experimental samples used to generate the data sets, essentially allowing a digital in situ of gene expression (Winter et al., 2007; Figure 2). Different filters can be selected to visualize expression in absolute terms, while a stimulus response can be visualized in the relative mode. The ability to monitor expression of two genes in the compare mode at high spatial resolution is useful when inferring regulatory relationships between genes of interest. Subcellular localization of a gene product can also be visualized using the Cell eFP Browser, whereby a confidence score for the localization of a gene product within each distinct subcellular compartment or region is calculated and displayed as a color scale (Winter et al., 2007). This tool will also be discussed in the proteomics section. Expression Angler is of great use when a researcher wants to identify genes that are similarly expressed with his or her gene of interest. These similarly expressed genes may be involved in the same biological process of the query gene or found within the same transcriptional regulatory module under the guilt-by-association paradigm. Taking an individual gene as bait, the user can set a Pearson correlation coefficient threshold to identify genes closely correlated or anticorrelated with that gene's expression pattern (Toufighi et al., 2005). An additional tool at the BAR is Promomer, which can identify statistically overrepresented *cis*-elements within the promoter region of a single gene or a list of genes, perhaps obtained from Expression Angler or Expression Browser (Toufighi et al., 2005).

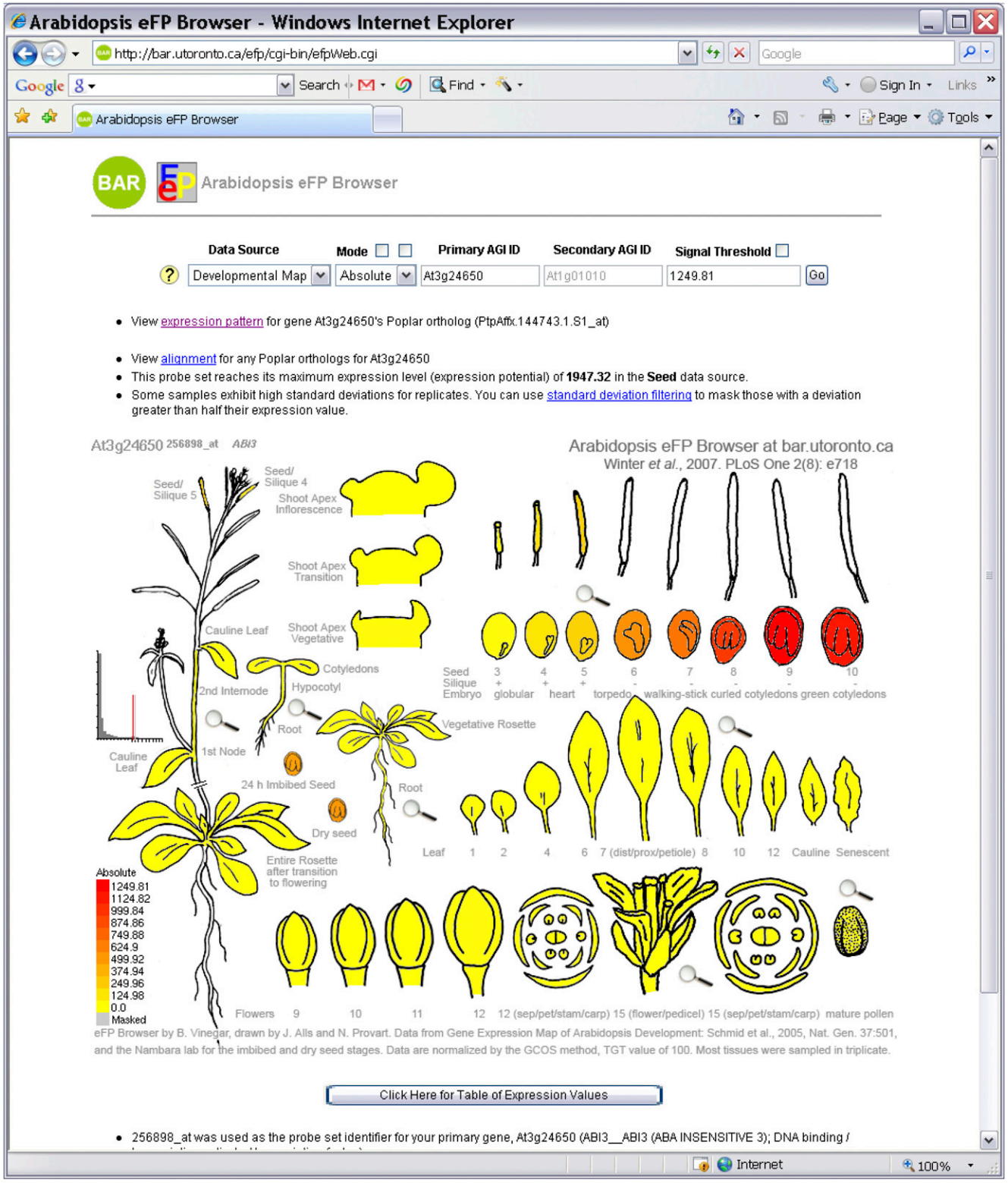

Figure 2. Exploring *Arabidopsis* Gene Expression Data with the eFP Browser (Winter et al., 2007).

Expression data for any one of ~24,000 genes are "painted" onto a pictographic representation of the samples that were used to generate the RNA for expression profiling. In this view, gene expression data are from the Schmid et al. (2005) Developmental Atlas and from the Nambara lab. Here, the expression level of *ABI3* (At3g24650) is seen to be highest toward the later stages of seed development, denoted by strong red coloration in the seed pictographs.

Bioexample 2: Hypothesis Generation and Validation Enabled by the eFP Browser

The usefulness of tools provided by the BAR is evident by the multitude of publications that have used this resource and its associated expression data. Within the last year, data visualized, in particular, by the eFP Browser have been used to both generate and validate hypotheses. One popular use of the eFP Browser is to aid in determining gene function by assessing expression within tissues. Visualized expression of the *MYBL2* transcription factor was used to guide tissue-specific characterization of the *MYBL2* response to light and control of flavonoid biosynthesis (Dubos et al., 2008), while eFP visualization of a sphingolipid Δ4-desaturase suggested that this desaturase is active within flowers, which was later confirmed in genetic analysis coupled with mass spectrometry (Michaelson et al., 2009). The eFP Browser can also be used to determine molecular mechanisms for gene function. An exploration of the tissue-specific roles of a holophytochrome, using tissue-specific promoters, suggested that *pBVR* specifically regulates far-red high irradiance responses in photosynthetic tissues (Warnasooriya and Montgomery, 2009). This far-red light specificity was determined not to depend on transcriptional control, but most likely requires additional regulatory mechanisms, since gene expression driven by the photosynthetic tissue reporter *CAB3* was similar in the different light conditions tested, as visualized by the eFP Browser (Warnasooriya and Montgomery, 2009). Finally, the eFP Browser was used as a tool to confirm and contrast in vivo expression experiments of photosynthesis-associated nuclear gene families in *Arabidopsis* (Sawchuk et al., 2008). Several examples of how another BAR tool, Expression Angler, has been used as a screen to identify novel genes involved in a specific pathway or process are also given in the Coexpression Tools section.

Genevestigator

Expression data analysis tools at the Genevestigator website allow users to answer many similar questions as with the BAR, specifically, how a gene or genes are expressed during the range of developmental stages and stimulus response conditions profiled by individual researchers and the AtGenExpress consortium for *Arabidopsis* as well as for other organisms. Of particular note, it is also possible to query gene expression within mutants using these tools (Zimmermann et al., 2005). Biological examples of how Genevestigator has been used to both generate and test hypotheses have been described by Grennan (2006). Genevestigator has recently been redesigned (Genevestigator V3), and its tools have been streamlined into four easy to use groupings (Hruz et al., 2008). Metaprofile analysis visualizes gene expression in heat map format across individual experiments or in the biological contexts of anatomy, development, stimulus, and mutation. The newly developed Biomarker Search tool can identify genes specifically expressed or repressed in a biological state (i.e., development, stimulus, or mutation). The Custom Bait feature of Expression Angler at the BAR offers similar functionality. The third toolset enables clustering analysis using two different methods, hierarchical clustering or biclustering, allowing identification of coexpressed and putatively coregulated groups of genes across a set of experimental conditions. Finally, the pathway projector tool incorporates manually verified reaction pathways and allows the user to overlay expression data onto these pathways. Local networks can be assembled by allowing the user to start with a single reaction or pathway and then extend it with neighboring reactions or pathways. All of these tools are integrated such that genes identified from one toolset can be incorporated into another. For example, gene expression across a group of developmental stages can be identified using the metaprofile analysis tool and then input into the clustering tool to generate hypotheses regarding which of these genes may be coregulated transcriptionally.

In cases where no probe set is present on the Affymetrix ATH1 microarray for one's gene of interest, or where the probe set on the ATH1 microarray hybridizes to transcripts from several genes, or where one's gene of interest can be associated with several gene models due to alternate splice forms, the At-TAX Web tool (gbrowse.weigelworld.org/cgi-bin/gbrowse/attax/) can be used to query whole-genome tiling array expression data for *Arabidopsis* development or stress responses (Laubinger et al., 2008). The MPSS website (http://mpss.udel.edu/at/) developed by Blake Meyers and colleagues at the University of Delaware also contains many short sequence reads of cDNAs generated by massively parallel signature sequencing and similar methods (Nakano et al., 2006). Because signature sequencing is not limited in its detection of transcripts to the corresponding probe being present on a microarray, expression data can be obtained for most *Arabidopsis* genes (rice data sets are also available). In addition, an RNA degradeome data set (Parallel Analysis of RNA Ends) has been loaded into the MPSS database (German et al., 2008).

Expression data atlases are available for a few other plant species, notably for poplar with PopGenIE at www.popgenie.db.umu.se (Sjödin et al., 2009) and with the BAR at BAR.utoronto.ca (Wilkins et al., 2008), and for *M. truncatula* with the Noble Foundation at bioinfo.noble.org/gene-atlas/ (Benedito et al., 2008). PLEXdb (plexdb.org) contains a barley expression atlas, in addition to selected expression data sets from several other agronomically important species, and from pathogens thereof (Shen et al., 2005).

The BAR, Genevestigator and other tools for exploring gene expression data are of great utility, but users must exercise caution when interpreting their results. Of prime importance is an awareness of raw expression values, normalization methods, and, in stimulus response experiments, expression levels within the control samples. Not being aware of these parameters can easily result in flawed interpretation of gene expression and in unsuccessful biological experiments. In the future, incorporation of much higher spatiotemporal resolution *Arabidopsis* root microarray expression data and of recently published high-resolution rice data is greatly needed. Incorporating these data or linking to tools that describe these data would make these queryable databases more comprehensive (Brady et al., 2007; Chaudhuri et al., 2008; Jiao et al., 2009).

Coexpression Tools

Often, genes that are coexpressed with one's gene of interest can provide an avenue for further exploration, particularly in

terms of association with a particular biological process. In addition to the aforementioned coexpression tools in Genevestigator and at the BAR, Aoki et al. (2007) have reviewed several other prominent coexpression tools: ACT, ATTEDII, and AthCoR@CSB.DB. CressExpress (Srinivasasainagendra et al., 2008) and GeneCAT (Mutwil et al., 2008) are more recent tools that are also useful for identifying coexpressed genes. Genes of unknown function within a list of genes that are highly coexpressed with a gene of interest may also be involved in the biological process of the query gene. There are several recent examples of coexpression being used as a primary screen to identify novel genes associated with a given biological process. Koo et al. (2006) performed a coexpression screen with genes that are coexpressed with known JA biosynthetic components to identify a key step in the jasmonic acid biosynthetic pathway in *Arabidopsis*. Hirai et al. (2007) pinpointed the transcription factors MYB28 and 29 as regulators of glucosinolate synthesis by combining coexpression across publicly available expression data sets, along with transcriptional analyses of sulfur-starved *Arabidopsis* plants. d'Erfurth et al. (2008) used Expression Angler to search for genes coexpressed with known meiotic genes and then phenotyped T-DNA mutants of 138 candidate genes. Chromosome spreads in two independent mutant alleles of At1g34355 (At *PS1*) revealed that these plants were polyploid, indicating a role in meiosis. Two other genes with meiotic function were identified in the same screen. As a final example, three new subunits of NAD(P)H dehydrogenase were similarly identified using coexpression and reverse genetics (Takabayashi et al., 2009).

Small RNA Databases

The importance of small RNAs in controlling many aspects of plant growth and development is one of the most exciting discoveries in plant biology in the past decade (Johnson and Sundaresan, 2007). Their role in such processes is certain to be revealed to be even more far reaching. For instance, it was recently shown that there is widespread inhibition of translation by miRNAs and small interfering RNAs (siRNAs, Brodersen et al., 2008), in addition to their more familiar roles in gene silencing and natural antisense. Several research groups have aimed to document all the small RNAs in *Arabidopsis* (Llave et al., 2002; Xie et al., 2005; Axtell et al., 2006; Rajagopalan et al., 2006; Fahlgren et al., 2007; Howell et al., 2007; Kasschau et al., 2007; Zilberman et al., 2007; Lister et al., 2008), which in turn have been collated into the *Arabidopsis* Small RNA Project (ASRP) Genome Browser at asrp.cgrb.oregonstate.edu (Gustafson et al., 2005). With this Genome View of the ASRP resource, it is possible to see if one's gene of interest is being targeted by a specific small RNA or contains elements that encode small RNAs that are being identified by the ASRP. Small RNAs identified from floral buds and immature flowers using deep sequencing technology can also be visualized in the UCSC *Arabidopsis* Genome Browser (Lister et al., 2008). For species other than *Arabidopsis*, the Cereal Small RNA Database (sundarlab.ucdavis.edu/smrnas/) contains large-scale data sets of maize and rice smRNA sequences generated by high-throughput pyrosequencing and have been mapped to the rice genome and available maize genome sequence (Johnson et al., 2007).

Bioexample 3: A Cell Type–Specific Nitrogen-Regulated Transcriptional Circuit That Mediates Developmental Plasticity

In elegant work examining the root cell type–specific response to nitrogen, Gifford et al. (2008) were able to elucidate a transcriptional circuit within the root pericycle involving a small RNA, miR167, and its negatively regulated target, *ARF8*, included in the ASRP that mediates developmental plasticity. In this work, the authors were able to show, using genetic and phenotypic analysis based on cell type–specific expression profiling data and knowledge of small RNA targets, that the expression level of *ARF8* was increased in response to nitrogen and that this was directly due to a nitrogen-stimulated decrease in miR167 production. This resulted in a high ratio of initiated lateral roots to emerged lateral roots under high nitrogen conditions. In nitrogen-depleted conditions, these initiated lateral roots can then emerge and explore the surrounding soil environment for nutrients.

Epigenetic Modifications

Transcription factor–mediated regulation of gene expression is only one component that determines the final level of gene expression. The marking of genes by the methylation of cytosines or by the methylation/acetylation of histones of the encompassing chromatin also can dramatically alter their level of expression. In the case of the *FLOWERING LOCUS C* (*FLC*) gene in *Arabidopsis*, dimethylation of Lys residues 9 and 27 on histone H3 of regions of the *FLC* locus serves to generate a memory of winter so that flowering does not occur until after winter is over (Bastow et al., 2004). Again, ingenious technologies, including chromatin immunoprecipitation (ChIP)/whole-genome tiling arrays and shotgun bisulphite sequencing, are allowing unprecedented insight into the epigenome of this species. Querying such data can allow researchers to determine the full complement of regulatory mechanisms that determine the expression of their gene of interest. For example, Zhang et al. (2007) performed ChIP/Chip using whole genome tiling arrays to examine the H3K27me3 patterns in *Arabidopsis*. Prior to this study, only seven genes had been shown to be H3K27me3 methylated, namely, *FLC*, *AGAMOUS*, *MEDEA*, *SHOOT MERISTEMLESS*, *PHERES1*, *FUSCA3*, and *AGAMOUS-LIKE19* (Zhang et al., 2007). Clearly, the above mentioned genes are developmentally important, and mutants unable to H3K27me3 methylate have severe developmental phenotypes. This whole-genome histone methylation study found that up to 4400 genes may be regulated by histone methylation. It is possible to search for the methylation pattern of one's gene of interest using the UCSC Genome Browser at epigenomics.mcdb.ucla.edu/H3K27m3/. In a similar manner, Zhang et al. (2006) used a whole-genome array approach to assay cytosine methylation status and found that genes that are cytosine methylated in their promoters are typically expressed in a tissue-specific manner, while those that are body methylated are expressed at higher

levels. These cytosine methylation marks can also be visualized in the TAIR Genome Browser. In another recent study, Cokus et al. (2008) perfected a method called BS-seq, combining the bisulphite treatment of genome DNA with Illumina short-read sequencing technology to generate a breathtaking base pair resolution map of cytosine methylation. These data have also been loaded into the UCSC Genome Browser at epigenomics.mcdb.ucla.edu/BS-Seq/. While it is certain to emerge that the epigenome is dynamic, these initial snapshots can provide insight into a researcher's genes of interest with respect to potential additional regulatory mechanisms.

Proteomics

While expression data can tell a researcher that a given gene is expressed (transcribed) under certain conditions or in certain tissues, whether or not the transcript is translated into a protein is another matter. Additionally, questions such as where the gene product might be localized within the cell, if there are any posttranslational modifications (e.g., phosphorylation), or if it interacts with other proteins, are important to answer to understand a given protein's function and activity. *Arabidopsis* proteomic data sets can be subdivided broadly into those attempting to quantify and document the proteome in different tissues and growth conditions, those that delimit subcellular localization, and those that tabulate protein–protein interactions.

A novel proteomic data set generated by linear trap quadrupole ion-trap mass spectrometry, which profiled protein presence in six organs and identified proteins for nearly 50% of annotated *Arabidopsis* gene models, is currently represented as a track on the Genome Browser at TAIR, and in the PRIDE BioMart (www.ebi.ac.uk/pride/prideMart.do) and is available in the queryable AtProteome server (fgcz-atproteome.unizh.ch). Many of these proteins were used to identify presumed organ-specific biomarkers based on approximate abundance values across different organs (Baerenfaller et al., 2008). Interestingly, some of these biomarkers were identified in a recent proteomic analysis of guard cells, demonstrating how important cell type resolution is in the generation of large-scale data sets (Zhao et al., 2008). Another large-scale proteomic data set that was acquired using two-dimensional liquid chromatographic fractionation followed by linear trap quadrupole ion-trap mass spectrometry on peptides from four different organs was published recently (Castellana et al., 2008). The authors also used TiO_2 to enrich for phosphopeptides, thus expanding our current data set with a sampling of the phosphoproteome. Interestingly, both of these approaches (Baerenfaller et al., 2008; Castellana et al., 2008) identified novel, previously unannotated proteins, enabling refinement of existing gene models, although neither obtained full proteome coverage. The Castellana et al. (2008) data set is deposited in the Tranche database (tranche.proteomecommons.org). Additionally, more than 6000 phosphopeptides from 10 published *Arabidopsis* studies are available from the PhosPhAt database at phosphat.mpimp-golm.mpg.de (Heazlewood et al., 2008). Including these data in TAIR would be of great use to the community. Few high-quality, quantitative proteomic data sets have been deposited in publicly available databases, with only one isotopic labeling mass spectrometry experiment deposited in 2005 to GEO (www.ncbi.nlm.nih.gov/projects/geo/) for *Arabidopsis* and two data sets at the Proteomics database (proteomics.Arabidopsis.info) set up by the Nottingham Arabidopsis Stock Centre (NASC). Several 2D gel data sets for a handful of developmental stages are available through the German federal government funded GABI (Genomanalyse im Biologischen System Pflanze) Project Primary Database website, www.gabipd.org (Riano-Pachon et al., 2009). This is in contrast with the thousands of gene expression data sets available for *Arabidopsis*. Jorrín et al. (2007) and Thelen and Peck (2007) both give good overviews of types of data sets from a variety of methodologies currently being applied to a number of different plant species and perhaps an idea of why, in spite of the many data sets generated, there isn't the equivalent of a Genevestigator or BAR eFP Browser tool available for them. Several reasons exist for this: first, proteomic data are much more complex than transcriptomic data, particularly in terms of the host of potential posttranslational modifications that could exist. Second, proteomics technologies are rapidly developing, for example, the iTRAQ method has only been in existence for a few years. Finally, a large number of proteomic experiments are required to obtain full proteome coverage for a sample of interest. Obtaining full proteome coverage over the large number of conditions available for *Arabidopsis* gene expression would be beyond the funding level of most plant research grants. That said, a MIAPE (for Minimum Information About a Proteomics Experiment) specification for proteomics data has been developed (Taylor et al., 2007) so that the aforementioned details (i.e., experimental metadata) are reported for published experiments.

Despite the lack of quantitative proteomics data sets in public repositories, such as GEO, several qualitative proteomics experiments using 2D gels followed up by mass spectrometric identification have been conducted to document the subcellular localization of proteins in *Arabidopsis*. The *Arabidopsis* Subcellular Database (SUBA) at www.plantenergy.uwa.edu.au/suba2/ (Heazlewood et al., 2007) has collated data from >1000 publications, documenting the subcellular localization of >6743 *Arabidopsis* proteins mainly based on mass spectrometry and green fluorescent protein fusion experimental data. The query interface allows the use of Boolean operators to look for overlaps in proteins identified in various published data sets, which is sometimes surprisingly low even for two proteomes from ostensibly the same subcellular compartment. Whether this is indicative of dynamic proteomes, difficulties in obtaining complete proteome coverage, or experimental error is unclear. Furthermore, predictions run with 10 common subcellular localization prediction programs have been applied to the entire *Arabidopsis* proteome with the results that most *Arabidopsis* proteins, if not documented, at least can be inferred to be in a certain compartment or compartments. Knowing the subcellular localization of a protein is vital for understanding its function. The BAR's Cell eFP Browser (Winter et al., 2007) at BAR.utoronto.ca displays SUBA's documented and predicted subcellular localizations in a pictographic manner, according to the confidence of the localization method, as described earlier.

No comprehensive data set exists for the *Arabidopsis* interactome, although several NSF 2010 projects are currently

underway to document this, for example, the interactions between membrane proteins and proteins in the "unknowneome." Two studies have attempted to predict interactions based on orthology to interacting proteins in other species (Geisler-Lee et al., 2007; Cui et al., 2008). In the case of the AtPID at atpid.biosino.org (Cui et al., 2008), the authors have also used coexpression matrices and protein domain co-occurrence to infer interaction. The 19,979 predicted interactions described in the Geisler-Lee et al. (2007) publication are available through the *Arabidopsis* Interactions Viewer (AIV) at the BAR (BAR.utoronto.ca), as well as 50,000 more from a more recent iteration of the approach using more organisms. The AIV also contains >2000 literature-documented, biochemically or genetically assayed interactions for *Arabidopsis*, some from studies with an individual protein and others from experiments conducted in a more high-throughput manner, such as those using protein microarrays to detect interactions between calmodulin-related proteins or mitogen-activated protein kinases and their protein targets (Popescu et al., 2007; 2009).

Bioexample 4: Putative Interactors in the SNARE-Syntaxin Pathway

Geisler-Lee et al. (2007) identified 20 putative interactors in the SNARE-syntaxin pathway using their predicted interactor approach. Previously, only eight interacting proteins had been described in the literature as components of this pathway, which is important for vesicle trafficking. Additionally, the authors used both the SUBA database of protein subcellular localization and coexpression analyses on AtGenExpress data sets to show that their predicted interactions likely are occurring in vivo based on the assumptions of colocalization and coexpression. Proteins involved in predicted interactions were found to be located more often than by chance in the same subcellular compartment, which is requisite for interaction. The genes encoding these predicted interactors also tended to be coexpressed spatially and temporally. Using the query interface of the AIV, it is possible to try to extend a list of genes based on predicted interactions. The resulting predicted interactors represent high-quality candidates for involvement in the biological system of interest, especially if they are also coexpressed and found to be in the same subcellular compartment.

Metabolomics

The output of the plant proteome is in part a huge diversity of small molecules, which is apparently many times more diverse than the small molecule component of mammalian proteomes (compare ~200,000 different small molecules in the plant kingdom space [Fiehn, 2002] to the ~6500 for humans as documented in the Human Metabolome database [www.hmdb.ca]). For a researcher, it is important to know what, if any, small molecule could be produced by a given gene product of interest (or if there are any small molecules that act upon it, which could be answered using BRENDA at www.brenda-enzymes.org if it is an enzyme) (Schomburg et al., 2002) or if a given stimulus/mutation causes an overall perturbation of the metabolome. Unfortunately, the scope of metabolomic experiments in plants is very small, with only a limited number of biological conditions examined to date and large gaps in our knowledge of biosynthetic pathways. This reflects the fact that many metabolomic methods are still in development, in part limited by the capabilities of current instrumentation, the development of a comprehensive set of library standards, and in the laborious annotation of as yet unidentified metabolites. Identification of these metabolites will complete our picture of biological processes occurring within plants by helping us to characterize metabolic pathways and their intermediates and signaling molecules more definitively.

The Golm Metabolome Database (csbdb.mpimp-golm.mpg.de/csbdb/gmd/gmd.html) contains several metabolomic experiments conducted on *Arabidopsis* plants grown, for example, under different light intensities (Kopka et al., 2005). It is possible to query the database for a given compound and to identify experiments for which the compound of interest was found to be higher or lower than a given threshold. MetNetDB (MetNetDB.org), out of Iowa State University, documents compounds in metabolic pathways and links these to gene products, in a manner similar to AraCyc at www.Arabidopsis.org/biocyc/ (Mueller et al., 2003), MapMan at www.gabipd.org/projects/MapMan/data.shtml (Thimm et al., 2004), KEGG Atlas at www.genome.jp/kegg/atlas/metabolism/ (Okuda et al., 2008), Reactome at reactome.org (Tsesmetzis et al., 2008), and other pathway databases. As metabolomic data sets become more prevalent, it would be highly desirable for GEO or some other larger database to serve as the primary repository for the raw data generated by these experiments. Other more specialized tools could then be developed based on subsets of data from the primary repository, a model that has worked very well for minimum information about a microarray experiment (MIAME)-compliant (Brazma et al., 2001) transcriptome data sets. Fiehn et al. (2005) operate BinBase at http://eros.fiehnlab.ucdavis.edu:8080/binbase-compound/, which documents >1000 small molecules from plants, and they and others are actively involved in the creation of the Metabolomics Standards Initiative to bring MIAME-like standards to metabolomics experiments (Fiehn et al., 2007).

Integrative Resources

The inclusion of each of these types of large-scale data sets in easy-to-use, queryable browsers is of great importance for hypothesis generation. In particular, genome browsers, like the TAIR Genome Browser, include multiple sources of data that users can integrate within their queries. This allows the user to identify genetic variation at the sequence level that may lead to alteration in regulation of gene expression or protein function across diverse accessions. Expression browsers, like the BAR and Genevestigator, which permit visualization or interrogation of gene expression at the anatomical level, or at the level of response to a stimulus, can be used to generate hypotheses about gene function. The identification of genes coexpressed with one's gene of interest, or clusters of genes that are coregulated, and the mining of these gene groups for functional

association via overrepresentation of Gene Ontologies allows for in silico prediction of gene function. Further mining of these lists for overrepresented upstream regulatory sequences can identify putative regulatory factors. Ultimately, however, integration of the types of data sets described here toward the generation of multilevel regulatory networks (multinetworks) for hypothesis generation is desirable. Furthermore, development of methods to query these networks in a statistical manner that also assesses and weighs the validity of the data sources is necessary.

Generation of a multinetwork that incorporates multiple sources of data in *Arabidopsis* has been accomplished in a queryable Web browser named VirtualPlant at VirtualPlant.org (Gutierrez et al., 2007a). This multinetwork incorporates data for metabolic pathways, known protein–protein, protein–DNA, miRNA–RNA, and predicted protein–protein and protein–DNA interactions (Gutierrez et al., 2007a). Resulting gene networks are visualized using Cytoscape, and regions of high connectivity can be identified using Antipole, a graph clustering algorithm (Ferro et al., 2003). The original VirtualPlant multinetwork contained 6176 gene nodes, 1459 metabolite nodes, and 230,900 edges (or interactions) between these nodes (Gutierrez et al., 2007a). This network has also recently been expanded to include bioinformatically identified protein–DNA interactions (Gutierrez et al., 2008). Subnetworks can be identified by querying multinetworks with a list of genes, often identified from gene expression analysis and statistically tested for significance. Functional annotations can also be overlaid upon identified subnetworks of the multinetwork to help infer subnetwork function. In one interesting approach (Thum et al., 2008), supernode networks were generated by collapsing genes from a subnetwork into a category according to both their metabolic pathways and the first two words of their gene annotation, although the statistical significance of these supernode annotations was not tested. The resulting size of the node is proportional to the number of genes annotated to that node (Thum et al., 2008). The VirtualPlant system has been used successfully to define gene networks in various signaling pathways as further described in Bioexample 5.

While VirtualPlant's integration of multiple data sources into a cohesive queryable system is an important advancement in our ability to make sense of and use large-scale data sets, attributing measures of confidence to an edge between two nodes, as defined by experimental evidence, would greatly improve accuracy in defining network interactions. For example, a predicted protein–DNA interaction should not be weighed as heavily as an experimentally verified protein–DNA interaction. Furthermore, for a set of experimentally verified interactions, interactions with multiple sources of experimental support should be given greater confidence than an interaction with a single source of experimental support. Attributing such measures of confidence is not a simple task, as many of these data sources are heterogeneous and require explicit knowledge of how each data set was obtained experimentally. Access to information, such as the statistical methods used to define a gene as expressed, a polymorphism as a deletion based on array hybridization signal, a promoter as marked by an epigenetic modification, a metabolite or protein as present and properly annotated, and the in planta relevance of interactions between plant proteins detected in yeast two hybrid assays should also be required when synthesizing multinetworks or when using data from these multinetworks.

Integration of diverse data types in a statistical framework to infer gene function or to identify gene or protein interactions has been accomplished for a wide variety of organisms, including yeast, mouse, and humans (Myers et al., 2005; Lee et al., 2007; Guan et al., 2008; Kim et al., 2008; Mostafavi et al., 2008; Ramani et al., 2008). Methods and guidelines to integrate and correlate such heterogeneous data have been described by Lee and Marcotte (2008) and provide good principles that should be taken into account in the plant community, especially as integration of multiple data sets has been shown to outperform individual functional genomics data sets in accuracy and coverage in hypothesis validation.

As an example, algorithms like GeneMANIA tackle this computationally complex problem at the level of the individual network using data from several levels (expression profiles, protein–protein interactions, subcellular localization, etc.). Functional prediction analyses are possible for several organisms, now including *Arabidopsis* (Mostafavi et al., 2008). The authors assign a weight to each network derived from a single data source that reflects its usefulness in predicting a given function of interest. To construct the final composite network, they then take the weighted average of the combined association networks. Furthermore, at the node level, GeneMANIA incorporates genes positively associated with a label from a particular network,

Bioexample 5: Elucidation of Gene Networks Using the VirtualPlant Multinetwork

The VirtualPlant network has been used to elucidate gene networks that act in response to light and carbon, carbon and nitrogen, and to organic nitrogen (Gutierrez et al., 2007a, 2007b; Thum et al., 2008). In these studies, microarray analysis was used to define a list of genes that responded combinatorially or individually to these stimuli. These lists were then used to query the VirtualPlant multinetwork and to define putative subgene networks. Subnetworks of high connectivity, or that contained specific types of regulatory connections, were then explored. A transcriptional regulatory subnetwork that acts in response to the assimilation of organic nitrogen (Glu/Gln) was defined by identifying transcription factors with the highest number of connections within the subnetwork. Of particular interest to the authors was the central clock oscillator gene, *CCA1*, and a golden 2-related transcription factor (*GLK1*). Both of these genes were predicted to activate expression of two genes involved in Gln metabolism/catabolism (*GLN1.3*/*GDH1*) and to repress a bZIP1 transcription factor that activates expression of a Gln-responsive gene (*ASN1*). Using a *CCA1* overexpressor line, *ASN1*, *GLN1.3*, *bZIP1*, and *GDH1* all showed altered expression patterns, genetically validating CCA1 as a regulator of these genes. Direct binding of CCA1 to *GLN1.3*, *GDH1*, and *bZIP1* was further confirmed by ChIP assays, validating this gene subnetwork. The influence of organic N on this subnetwork was tested by monitoring the effects of N on the oscillatory expression of *CCA1*. Gln in particular was shown to shorten the oscillatory period, thereby demonstrating that organic nitrogen status feeds into the circadian clock via CCA1 and regulates N metabolism downstream.

genes that are unlabeled, and genes that are negatively labeled. This method has proven more accurate in prediction of GO category association than leading methods on mouse and yeast functional data, using the area under the curve for the resulting receiver operating characteristic curves (Pena-Castillo et al., 2008). For the mouse data, this is primarily due to their inclusion of genes that are negatively labeled. GeneMANIA is also available on a Web server for easy access (see Table 1). The application or development of such algorithms to available plant large-scale data sets is greatly needed. In addition, the concept of a competition for critical assessment of *Arabidopsis* gene function prediction, similar to that held for mouse (Pena-Castillo et al., 2008), in which several groups submit their best predictions on a benchmark data set assembled by organizers, might be attractive, especially considering some of the novel data types available in this species.

Future Directions

Comparisons across species are often key to understanding a biological process. To make accurate cross-species comparisons, it is necessary to have a vocabulary representing the similarity of form and function in the species under consideration. For this reason, the Plant Ontology resource was developed based on anatomical and functional features of *Arabidopsis*, rice, and maize, with others being added to describe other crop species (plantontology.org). Ontology terms include those describing tissues and cell types, organs and organ systems, and those denoting particular stages, such as senescence or germination. To permit the incorporation of data from as yet unsequenced genomes, a standardized staging and developmental state for each plant organ was developed (Jaiswal et al., 2005; Ilic et al., 2007). Careful annotation of the tissue samples using the Plant Ontology system would make it easy to query, for example, the response of orthologous genes in similar tissues in related or more distant species. In a similar manner, a recently funded National Institutes of Health Genome Research Resource Grant (P41) to establish a Pathway Commons to facilitate the exchange, integration, and distribution of biological pathway information will maintain and extend the BioPAX exchange language for biological pathways and develop improved software for querying these.

The iPlant Collaborative (iplantcollaborative.org) recently decided on which Grand Challenge proposals to assist Plant Cyberinfrastructure will be financially supported. These are Assembling the Tree of Life for the Plant Sciences, which is focused on the design and creation of a phylogenetic cyberinfrastructure, and Cyberinfrastructural Support for the Genetic and Ecophysiological Decipherment of Plant Phenological Control in Complex and Changing Environments.

A looming challenge, presumably in part to be addressed by the second iPlant Collaborative Grand Challenge listed above, presents itself with next-generation sequencing initiatives, which have revolutionized genome sequencing abilities in terms of time and cost. Within the next 10 years, thousands of plant genome sequences will be released, using a variety of different sequencing platforms, including the Roche 454 pyrosequencing system, the Illumina Genome Analyzer, and the Applied Biosystems SOLiD system. Plant research is benefitting from these technologies, with several *Arabidopsis* accessions recently sequenced and many more plant genomes in the sequencing pipeline (Ossowski et al., 2008). Once these large data sets are publicly released, the next challenge is in how to deal with the resulting bioinformatic bottlenecks. In particular, base calling software, methods to score sequence quality, and alignment software can all differ depending on both the sequencing platform and the researcher's personal choice and should be accounted for in sequence browsers that publicly display these data. Available bioinformatic software is well described by Shendure and Ji (2008). Community accepted guidelines for compiling these and other metadata associated with these data sets, in a manner similar to that of MIAME guidelines (Brazma et al., 2001), will be extremely important in the future. A draft version of this guideline, termed MINSEQE (minimum information about a high-throughput sequencing experiment) has been proposed (www.mged.org/minseqe/). NCBI has established a short read archive (SRA) (http://0-www.ncbi.nlm.nih.gov.lib1.npue.edu.tw/Traces/sra/) that accepts next-generation sequencing data from a variety of platforms and includes data from de novo sequencing experiments, resequencing experiments, structural variation discovery, and SNP calling experiments. The SRA tracks metadata associated with each experiment and should help improve database efficiency by normalizing data structures. GEO accepts next-generation sequencing data from mRNA sequencing, ChIP sequencing, bisulfite sequencing, and small RNA discovery and profiling experiments (www.ncbi.nlm.nih.gov/projects/geo/info/seq.html). Metadata are also associated with these submissions. A remaining challenge is in displaying this wealth of sequence data in a user-queryable, Web interface format in a manner that allows the user to extract biologically meaningful information from these data sets.

Clearly, the trend toward generating more and more data, especially sequence and expression data, will necessitate the development of new computational tools for viewing, querying, and analyzing such data. How will the average "wet lab" scientist be able to use data from 1001 genomes, let alone view them? Interestingly, it would seem that the generation of such data will lead to the reunification of the ecology and evolutionary, and cell and molecular fields of plant biology.

In the meantime, so-called SOAP (Simple Object Access Protocol) services, such as various BioMOBY resources (Wilkinson and Links, 2002), are being developed by plant bioinformatic groups worldwide. These services promise to allow databases and Web-based tools to "talk" to one another, thereby automating specific aspects of creatively thought out analyses now becoming possible with ever more published large-scale data sets. For the average plant biologist, there is already a wealth of information available with existing Web-based tools, and he or she would be wise to embrace the computer as the "new molecular biology."

ACKNOWLEDGMENTS

We thank Wolfgang Busch, Geoff Fucile, Anjali Iyer-Pascuzzi, Julie Kang, Terri Long, David Orlando, Mallorie Taylor, and Jaimie VanNorman for critical review of the manuscript. We also thank Luca Comai for

insightful discussions on next-generation sequencing. Finally, our sincere thanks to both the reviewers and the editors for their comments, suggestions, and insight. N.J.P. is funded by a research grant from National Sciences and Engineering Council of Canada.

Received February 1, 2009; revised April 3, 2009; accepted April 12, 2009; published April 28, 2009.

REFERENCES

Alonso, J.M., et al. (2003). Genome-wide insertional mutagenesis of *Arabidopsis thaliana.* Science **301:** 653–657.

Aoki, K., Ogata, Y., and Shibata, D. (2007). Approaches for extracting practical information from gene co-expression networks in plant biology. Plant Cell Physiol. **48:** 381–390.

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

Axtell, M.J., Jan, C., Rajagopalan, R., and Bartel, D.P. (2006). A two-hit trigger for siRNA biogenesis in plants. Cell **127:** 565–577.

Baerenfaller, K., Grossmann, J., Grobei, M.A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., Zimmermann, P., Grossniklaus, U., Gruissem, W., and Baginsky, S. (2008). Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. Science **320:** 938–941.

Bastow, R., Mylne, J.S., Lister, C., Lippman, Z., Martienssen, R.A., and Dean, C. (2004). Vernalization requires epigenetic silencing of FLC by histone methylation. Nature **427:** 164–167.

Benedito, V.A., et al. (2008). A gene expression atlas of the model legume *Medicago truncatula.* Plant J. **55:** 504–513.

Berger, M.F., et al. (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell **133:** 1266–1276.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat. Biotechnol. **24:** 1429–1435.

Brady, S.M., Orlando, D.A., Lee, J.-Y., Wang, J.Y., Koch, J., Dinneny, J.R., Mace, D., Ohler, U., and Benfey, P.N. (2007). A high-resolution root spatiotemporal map reveals dominant expression patterns. Science **318:** 801–806.

Brazma, A., et al. (2001). Minimum information about a microarray experiment (MIAME) - Toward standards for microarray data. Nat. Genet. **29:** 365–371.

Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y.Y., Sieburth, L., and Voinnet, O. (2008). Widespread translational inhibition by plant miRNAs and siRNAs. Science **320:** 1185–1190.

Castellana, N.E., Payne, S.H., Shen, Z., Stanke, M., Bafna, V., and Briggs, S.P. (2008). Discovery and revision of *Arabidopsis* genes by proteogenomics. Proc. Natl. Acad. Sci. USA **105:** 21034–21038.

Chaudhuri, B., Hörmann, F., Lalonde, F., Brady, S.M., Orlando, D.A., Benfey, P., and Frommer, W.B. (2008). Protonophore- and pH-insensitive glucose and sucrose accumulation detected by FRET nanosensors in Arabidopsis root tips. Plant J. **56:** 948–962.

Chory, J., et al. (2000). National Science Foundation-sponsored workshop report: "The 2010 Project" functional genomics and the virtual plant. A blueprint for understanding how plants are built and how to improve them. Plant Physiol. **123:** 423–426.

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

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., and Jacobsen, S.E. (2008). Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. Nature **452:** 215–219.

Coruzzi, G., Gutierrez, R., Shasha, D., Katari, M., Gifford, M., Birnbaum, K., and Poultney, L. (2006). A systems approach to nitrogen networks and the "VirtualPlant". Dev. Biol. **295:** 327

Cui, J., Li, P., Li, G., Xu, F., Zhao, C., Li, Y., Yang, Z., Wang, G., Yu, Q., Li, Y., and Shi, T. (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., and Grotewold, E. (2003). AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC Bioinformatics **4:** 25.

d'Erfurth, I., Jolivet, S., Froger, N., Catrice, O., Novatchkova, M., Simon, M., Jenczewski, E., and Mercier, R. (2008). Mutations in *AtPS1* (*Arabidopsis thaliana Parallel Spindle 1*) lead to the production of diploid pollen grains. PLoS Genet. **4:** e1000274.

Dubos, C., Le Gourrierec, J., Baudry, A., Huep, G., Lanet, E., Debeaujon, I., Routaboul, J.-M., Alboresi, A., Weisshaar, B., and Lepiniec, L. (2008). MYBL2 is a new regulator of flavonoid biosynthesis in *Arabidopsis thaliana.* Plant J. **55:** 940–953.

Dunkley, T.P.J., et al. (2006). Mapping the *Arabidopsis* organelle proteome. Proc. Natl. Acad. Sci. USA **103:** 6518–6523.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. **30:** 207–210.

Fahlgren, N., Howell, M.D., Kasschau, K.D., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., Law, T.F., Grant, S.R., Dangl, J.L., and Carrington, J.C. (2007). High-throughput sequencing of Arabidopsis microRNAs: Evidence for frequent birth and death of miRNA genes. PLoS One **2:** e219.

Ferro, A., Pigola, G., Pulvirenti, A., and Shasha, D. (2003). Fast clustering and minimum weight matching algorithms for very large mobile backbone wireless networks. Int. J. Found. Comput. Sci. **14:** 223–236.

Fiehn, O. (2002). Metabolomics – The link between genotypes and phenotypes. Plant Mol. Biol. **48:** 155–171.

Fiehn, O., Sumner, L., Rhee, S., Ward, J., Dickerson, J., Lange, B., Lane, G., Roessner, U., Last, R., and Nikolau, B. (2007). Minimum reporting standards for plant biology context information in metabolomic studies. Metabolomics **3:** 195–201.

Fiehn, O., Wohlgemuth, G., and Scholz, M. (2005). Setup and annotation of metabolomic experiments by integrating biological and mass spectrometric metadata. In Data Integration in the Life Sciences, S. Istrail, P. Pevzner, and M. Waterman, eds (Berlin: Springer), pp. 224–239.

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M., and Dubchak, I. (2004). VISTA: Computational tools for comparative genomics. Nucleic Acids Res. **32:** W273–279.

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.

Gentleman, R., et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol. **5:** R80.

German, M.A., et al. (2008). Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. Nat. Biotechnol. **26:** 941–946.

Gifford, M.L., Dean, A., Gutierrez, R.A., Coruzzi, G.M., and Birnbaum, K.D. (2008). Cell-specific nitrogen responses mediate developmental plasticity. Proc. Natl. Acad. Sci. USA **105:** 803–808.

Goda, H., et al. (2008). The AtGenExpress hormone- and chemical-treatment data set: Experimental design, data evaluation, model data analysis, and data access. Plant J. **130:** 1319–1334.

Grennan, A.K. (2006). Genevestigator: Facilitating web-based expression analysis. Plant Physiol. **141:** 1164–1166.

Guan, Y., Myers, C.L., Lu, R., Lemischka, I.R., Bult, C.J., and Troyanskaya, O.G. (2008). A genomewide functional network for the laboratory mouse. PLOS Comput. Biol. **4:** e1000165.

Gustafson, A.M., Allen, E., Givan, S., Smith, D., Carrington, J.C., and Kasschau, K.D. (2005). ASRP: The *Arabidopsis* Small RNA Project database. Nucleic Acids Res. **33:** D637–D640.

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

Gutierrez, R.A., Gifford, M.L., Poultney, C., Wang, R., Shasha, D.E., Coruzzi, G.M., and Crawford, N.M. (2007b). Insights into the genomic nitrate response using genetics and the Sungear Software System. J. Exp. Bot. **58:** 2359–2367.

Gutierrez, R.A., Stokes, T.L., Thum, K., Xu, X., Obertello, M., Katari, M.S., Tanurdzic, M., Dean, A., Nero, D.C., McClung, C.R., and Coruzzi, G.M. (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.

Hanikenne, M., Talke, I.N., Haydon, M.J., Lanz, C., Nolte, A., Motte, P., Kroymann, J., Weigel, D., and Kramer, U. (2008). Evolution of metal hyperaccumulation required cis-regulatory changes and triplication of HMA4. Nature **453:** 391–395.

Heazlewood, J.L., Durek, P., Hummel, J., Selbig, J., Weckwerth, W., Walther, D., and Schulze, W.X. (2008). PhosPhAt: A database of phosphorylation sites in *Arabidopsis thaliana* and a plant-specific phosphorylation site predictor. Nucleic Acids Res. **36:** D1015–D1021.

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.

Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T. (1999). Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res. **27:** 297–300.

Hirai, M.Y., et al. (2007). Omics-based identification of *Arabidopsis* Myb transcription factors regulating aliphatic glucosinolate biosynthesis. Proc. Natl. Acad. Sci. USA **104:** 6478–6483.

Howell, M.D., Fahlgren, N., Chapman, E.J., Cumbie, J.S., Sullivan, C.M., Givan, S.A., Kasschau, K.D., and Carrington, J.C. (2007). Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in *Arabidopsis* reveals dependency on miRNA- and tasiRNA-directed targeting. Plant Cell **19:** 926–942.

Hruz, T., Laule, O., Szabo, G., Wessendorp, F., Bleuler, S., Oertle, L., Widmayer, P., Gruissem, W., and Zimmermann, P. (2008). Genevestigator V3: A reference expression database for the meta-analysis of transcriptomes. Adv. Bioinform. **420747**.

Iida, K., Seki, M., Sakurai, T., Satou, M., Akiyama, K., Toyoda, T., Konagaya, A., and Shinozaki, K. (2004). Genome-wide analysis of alternative pre-mRNA splicing in *Arabidopsis thaliana* based on full-length cDNA sequences. Nucleic Acids Res. **32:** 5096–5103.

Ilic, K., et al. (2007). The Plant Structure Ontology, a unified vocabulary of anatomy and morphology of a flowering plant. Plant Physiol. **143:** 587–599.

Jaiswal, P., et al. (2005). Plant Ontology (PO): A controlled vocabulary of plant structures and growth stages. Comp. Funct. Genomics **6:** 388–397.

Jiao, Y., et al. (2009). A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. Nat. Genet. **41:** 258–263.

Johnson, C., Bowman, L., Adai, A.T., Vance, V., and Sundaresan, V. (2007). CSRDB: A small RNA integrated database and browser resource for cereals. Nucleic Acids Res. **35:** D829–D833.

Johnson, C., and Sundaresan, V. (2007). Regulatory small RNAs in plants. EXS **97:** 99–113.

Jorrín, J., Maldonado, A., and Castillejo, M. (2007). Plant proteome analysis: A 2006 update. Proteomics **7:** 2947–2962.

Kasschau, K.D., Fahlgren, N., Chapman, E.J., Sullivan, C.M., Cumbie, J.S., Givan, S.A., and Carrington, J.C. (2007). Genome-wide profiling and analysis of *Arabidopsis* siRNAs. PLoS Biol. **5:** e57.

Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angelo, C., Bornberg-Bauer, E., Kudla, J., and Harter, K. (2007). The AtGenExpress global stress expression data set: Protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. Plant J. **50:** 347–363.

Kim, W., Krumpelman, C., and Marcotte, E. (2008). Inferring mouse gene functions from genomic-scale data using a combined functional network/classification strategy. Genome Biol. **9:** S5.

Koo, A.J.K., Chung, H.S., Kobayashi, Y., and Howe, G.A. (2006). Identification of a peroxisomal acyl-activating enzyme involved in the biosynthesis of jasmonic acid in *Arabidopsis.* J. Biol. Chem. **281:** 33511–33520.

Kopka, J., et al. (2005). GMD@CSB.DB: The Golm Metabolome Database. Bioinformatics **21:** 1635–1638.

Kubis, S., Baldwin, A., Patel, R., Razzaq, A., Dupree, P., Lilley, K., Kurth, J., Leister, D., and Jarvis, P. (2003). The *Arabidopsis ppi1* mutant is specifically defective in the expression, chloroplast import, and accumulation of photosynthetic proteins. Plant Cell **15:** 1859–1871.

Laubinger, S., Zeller, G., Henz, S., Sachsenberg, T., Widmer, C., Naouar, N., Vuylsteke, M., Scholkopf, B., Ratsch, G., and Weigel, D. (2008). At-TAX: A whole genome tiling array resource for developmental expression analysis and transcript identification in *Arabidopsis thaliana.* Genome Biol. **9:** R112.

Lee, I., Li, Z., and Marcotte, E.M. (2007). An improved, bias-reduced probabilistic functional gene network of baker's yeast, *Saccharomyces cerevisiae.* PLoS One **2:** e988.

Lee, I., and Marcotte, E.M. (2008). Integrating functional genomics data. Methods Mol. Biol. **453:** 267–278.

Lescot, M., Dehais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., Rouze, P., and Rombauts, S. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res. **30:** 325–327.

Liang, C., et al. (2008). Gramene: A growing plant comparative genomics resource. Nucleic Acids Res. **36:** D947–D953.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H., and Ecker, J.R. (2008). Highly integrated single-base resolution maps of the epigenome in *Arabidopsis.* Cell **133:** 523–536.

Llave, C., Kasschau, K.D., Rector, M.A., and Carrington, J.C. (2002). Endogenous and silencing-associated small RNAs in plants. Plant Cell **14:** 1605–1619.

Michaelson, L.V., Zauner, S., Markham, J.E., Haslam, R.P., Desikan, R., Mugford, S., Albrecht, S., Warnecke, D., Sperling, P., Heinz, E., and Napier, J.A. (2009). Functional characterization of a higher plant sphingolipid Δ4-desaturase: Defining the role of sphingosine and sphingosine-1-phosphate in *Arabidopsis.* Plant Physiol. **149:** 487–498.

Mitchell-Olds, T., and Schmitt, J. (2006). Genetic mechanisms and evolutionary significance of natural variation in *Arabidopsis.* Nature **441:** 947–952.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., and Morris, Q. (2008). GeneMANIA: A real-time multiple association network integration algorithm for predicting gene function. Genome Biol. **9:** S4.

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

Mutwil, M., Obro, J., Willats, W.G.T., and Persson, S. (2008). GeneCAT: Novel webtools that combine BLAST and co-expression analyses. Nucleic Acids Res. **36:** W320–326.

Myers, C., Robson, D., Wible, A., Hibbs, M., Chiriac, C., Theesfeld, C., Dolinski, K., and Troyanskaya, O. (2005). Discovery of biological networks from diverse functional genomic data. Genome Biol. **6:** R114.

Nakano, M., Nobuta, K., Vemaraju, K., Tej, S.S., Skogen, J.W., and Meyers, B.C. (2006). Plant MPSS databases: Signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Res. **34:** D731–D735.

O'Connor, T.R., Dyreson, C., and Wyrick, J.J. (2005). Athena: A resource for rapid visualization and systematic analysis of Arabidopsis promoter sequences. Bioinformatics **21:** 4411–4413.

Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S., and Kanehisa, M. (2008). KEGG Atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res. **36:** W423–426.

Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N., and Weigel, D. (2008). Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res. **18:** 2023–2033.

Pena-Castillo, L., et al. (2008). A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence. Genome Biol. **9:** S2.

Popescu, S.C., Popescu, G.V., Bachan, S., Zhang, Z., Gerstein, M., Snyder, M., and Dinesh-Kumar, S.P. (2009). MAPK target networks in *Arabidopsis thaliana* revealed using functional protein microarrays. Genes Dev. **23:** 80–92.

Popescu, S.C., Popescu, G.V., Bachan, S., Zhang, Z., Seay, M., Gerstein, M., Snyder, M., and Dinesh-Kumar, S.P. (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.

Quackenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. (2000). The TIGR Gene Indices: Reconstruction and representation of expressed gene sequences. Nucleic Acids Res. **28:** 141–145.

Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D.P. (2006). A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana.* Genes Dev. **20:** 3407–3425.

Ramani, A.K., Li, Z., Hart, G.T., Carlson, M.W., Boutz, D.R., and Marcotte, E.M. (2008). A map of human protein interactions derived from co-expression of human mRNAs and their orthologs. Mol. Syst. Biol. **4:** 180.

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

Riano-Pachon, D.M., Nagel, A., Neigenfind, J., Wagner, R., Basekow, R., Weber, E., Mueller-Roeber, B., Diehl, S., and Kersten, B. (2009). GabiPD: The GABI primary database - a plant integrative 'omics' database. Nucleic Acids Res. **37:** D954–D959.

Sawchuk, M.G., Donner, T.J., Head, P., and Scarpella, E. (2008). Unique and overlapping expression patterns among members of photosynthesis-associated nuclear gene families in *Arabidopsis.* Plant Physiol. **148:** 1908–1924.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D., and Lohmann, J.U. (2005). A gene expression map of *Arabidopsis thaliana* development. Nat. Genet. **37:** 501–506.

Schomburg, I., Chang, A., Hofmann, O., Ebeling, C., Ehrentreich, F., and Schomburg, D. (2002). BRENDA: A resource for enzyme data and metabolic information. Trends Biochem. Sci. **27:** 54–56.

Seki, M., et al. (2002). Functional annotation of a full-length *Arabidopsis* cDNA collection. Science **296:** 141–145.

Seki, M., et al. (2004). RIKEN *Arabidopsis* full-length (RAFL) cDNA and its applications for expression profiling under abiotic stress conditions. J. Exp. Bot. **55:** 213–223.

Shen, L., Gong, J., Caldo, R.A., Nettleton, D., Cook, D., Wise, R.P., and Dickerson, J.A. (2005). BarleyBase - An expression profiling database for plant genomics. Nucleic Acids Res. **33:** D614–D618.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat. Biotechnol. **26:** 1135–1145.

Sjödin, A., Street, N.R., Sandberg, G., Gustafsson, P., and Jansson, S. (2009). The *Populus* Genome Integrative Explorer (PopGenIE): A new resource for exploring the *Populus* genome. New Phytol. http://dx.doi.org/10.1111/j.1469-8137.2009.02807.x/.

Srinivasasainagendra, V., Page, G.P., Mehta, T., Coulibaly, I., and Loraine, A.E. (2008). CressExpress: A tool for large-scale mining of expression data from *Arabidopsis.* Plant Physiol. **147:** 1004–1016.

Steffens, N.O., Galuschka, C., Schindler, M., Bulow, L., and Hehl, R. (2005). AthaMap web tools for database-assisted identification of combinatorial *cis*-regulatory elements and the display of highly conserved transcription factor binding sites in *Arabidopsis thaliana.* Nucleic Acids Res. **33:** W397–402.

Swarbreck, D., et al. (2008). The *Arabidopsis* Information Resource (TAIR): Gene structure and function annotation. Nucleic Acids Res. **36:** D1009–D1014.

Takabayashi, A., Ishikawa, N., Obayashi, T., Ishida, S., Obokata, J., Endo, T., and Sato, F. (2009). Three novel subunits of *Arabidopsis* chloroplastic NAD(P)H dehydrogenase identified by bioinformatic and reverse genetic approaches. Plant J. **57:** 207–219.

Taylor, C.F., et al. (2007). The minimum information about a proteomics experiment (MIAPE). Nat. Biotechnol. **25:** 887–893.

Thelen, J.J., and Peck, S.C. (2007). Quantitative proteomics in plants: Choices in abundance. Plant Cell **19:** 3339–3346.

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.

Thum, K., Shin, M., Gutierrez, R., Mukherjee, I., Katari, M., Nero, D., Shasha, D., and Coruzzi, G. (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.

Toufighi, K., Brady, M., Austin, R., Ly, E., and Provart, N. (2005). The Botany Array Resource: e-northerns, expression angling, and promoter analyses. Plant J. **43:** 153–163.

Tsesmetzis, N., et al. (2008). *Arabidopsis* Reactome: A foundation knowledgebase for plant systems biology. Plant Cell **20:** 1426–1436.

Ware, D. (2007). Gramene: A resource for comparative grass genomics. Methods Mol. Biol. **406:** 315–329.

Warnasooriya, S.N., and Montgomery, B.L. (2009). Detection of spatial-specific phytochrome responses using targeted expression of biliverdin reductase in *Arabidopsis.* Plant Physiol. **149:** 424–433.

Wilkins, O., Nahal, H., Foong, J., Provart, N.J., and Campbell, M.M. (2008). Expansion and diversification of the *Populus* R2R3-MYB family of transcription factors. Plant Physiol. **149:** 981–993.

Wilkinson, M.D., and Links, M. (2002). BioMOBY: An open source biological web services proposal. Brief. Bioinform. **3:** 331–341.

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V., and Provart, N.J. (2007). An 'electronic fluorescent pictograph' browser for exploring and analyzing large-scale biological data sets. PLoS One **2:** e718.

Xie, Z., Allen, E., Fahlgren, N., Calamar, A., Givan, S.A., and

Carrington, J.C. (2005). Expression of *Arabidopsis* miRNA genes. Plant Physiol. **138:** 2145–2154.

Yamamoto, Y.Y., and Obokata, J. (2008). ppdb: A plant promoter database. Nucleic Acids Res. **36:** D977–D981.

Zeller, G., Clark, R.M., Schneeberger, K., Bohlen, A., Weigel, D., and Ratsch, G. (2008). Detecting polymorphic regions in *Arabidopsis thaliana* with resequencing microarrays. Genome Res. **18:** 918–929.

Zhang, X., Clarenz, O., Cokus, S., Bernatavichute, Y.V., Pellegrini, M., Goodrich, J., and Jacobsen, S.E. (2007). Whole-genome analysis of histone H3 lysine 27 trimethylation in *Arabidopsis.* PLoS Biol. **5:** e129.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W.L., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S. E., and Ecker, J.R. (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis.* Cell **126:** 1189–1201.

Zhao, Z., Zhang, W., Stanley, B.A., and Assmann, S.M. (2008). Functional proteomics of *Arabidopsis thaliana* guard cells uncovers new stomatal signaling pathways. Plant Cell **20:** 3210–3226.

Zhu, T., and Wang, X. (2000). Large-scale profiling of the *Arabidopsis* transcriptome. Plant Physiol. **124:** 1472–1476.

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T., and Henikoff, S. (2007). Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. Nat. Genet. **39:** 61–69.

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.