The *Arabidopsis* genome: A foundation for plant research

Michael Bevan and Sean Walsh

Genome Res. 2005 15: 1632-1642
Access the most recent version at doi:10.1101/gr.3723405

Perspective

The *Arabidopsis* genome: A foundation for plant research

Michael Bevan[1,3] and Sean Walsh[2]

[1]*Cell and Developmental Biology Department,* [2]*Computational Biology Department, John Innes Centre, Norwich NR4 7UJ, United Kingdom*

The sequence of the first plant genome was completed and published at the end of 2000. This spawned a series of large-scale projects aimed at discovering the functions of the 25,000+ genes identified in *Arabidopsis thaliana* (*Arabidopsis*). This review summarizes progress made in the past five years and speculates about future developments in *Arabidopsis* research and its implications for crop science. The provision of large populations of gene disruption lines to the research community has greatly accelerated the impact of genomics on many areas of plant science. The tools and community organization required for plant integrative and systems biology approaches are now ready to accomplish the next big step in plant biology—the integration of knowledge and modeling of biological processes. In the future, plant science will continue to be enriched by the alignment of high-quality basic research (generally conducted in *Arabidopsis*), with strategic objectives in crop plants. The sequence and analysis of an increasing number of crop plant genomes enhance this alignment and provide new insights into genome evolution and crop plant domestication.

[The following individuals kindly provided reagents, samples, or unpublished information as indicated in the paper: P. Prusinkiewicz, E. Coen.]

Arabidopsis thaliana was the first plant, and the third multicellular organism after *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium 1998) and *Drosophila melanogaster* (Adams et al. 2000), to be completely sequenced (The *Arabidopsis* Genome Initiative 2000). At the time, it was claimed that the *Arabidopsis* genome sequence "… creates the potential for direct and efficient access to a much deeper understanding of plant development and environmental responses, and permits the structure and dynamics of plant genomes to be assessed and understood." Five years on, how justified was this claim? Furthermore, a vision for the *Arabidopsis* research community was articulated based on the promise of the genome sequence. A noteworthy aspiration was to "determine the function of all *Arabidopsis* genes by 2010." What progress has been made toward this goal? This review takes a broad and necessarily shallow view of progress in *Arabidopsis* research and relates this to work in other reference organisms. Our analysis suggests that much of the extraordinary progress made in the past five years has drawn on the genome sequence, and shows that it has had a catalytic effect on the research community and on how plant science is conducted. However, the explosion of data has created unanticipated problems that the *Arabidopsis* and plant science community must address if it is to take successfully to the path of integrative biology.

Progress in sequencing, reassembly, and annotation of the genome

Since systematic sequencing was completed in late 2000, the genome sequence has undergone several rounds of reassembly, hole patching, and extension into unsequenced regions. The sequenced and analyzed regions of the genome cover ~119 Mb (million base pairs) of genome sequence. The most impressive accomplishment has been to extend BAC and YAC contigs further into pericentromeric regions (Hosouchi et al. 2002). The size of the remaining gap in each chromosome, estimated using gel electrophoresis, varied between 4 Mb and 9 Mb, yielding an overall genome size of ~146 Mb. Several new genes were also identified in the pericentromeric heterochromatin. These regions are among the most comprehensively described (Hall et al. 2002) and are a key resource for understanding and using centromere functions, for example, to make minichromosomes, for understanding the biological functions of repeat sequences and how they originate and are maintained.

The initial set of gene models was generated by a combination of optimized ab initio gene-finding algorithms and supporting data such as EST and cDNA sequence. This was carried out in several locations (for maximum speed) and inevitably discrepancies occurred. Subsequently more systematic rounds of annotation incorporated a large number of full-length (FL) cDNA sequences (notably the RIKEN/SALK and Ceres resources) such that 16,000 of the 29,000 predicted genes are supported by experimental evidence (Wortman et al. 2003). Comparison of *Arabidopsis* sequences with genomic sequence from the closely related *Brassica oleracea* (Chinese cabbage) identified regions of high similarity that either identified putative new genes or extended existing gene models. About 30% of these new genes encoded a transcript. About 25% of the originally predicted genes had no supporting evidence such as an EST match or reasonable similarity of their putative peptide sequence to any other protein (the hypothetical genes), and consequently their biological relevance remains doubtful. Nevertheless, systematic analysis of this class of genes on Chromosome 2 revealed that the majority of these were expressed (Xiao et al. 2002). Other analyses of EST matches to the genome sequence found several hundred matches to putative noncoding regions and other noncanonical gene-like entities (Riano-Pachon et al. 2005).

The canonical genome sequence and annotation is cur-

[3]Corresponding author.
E-mail michael.bevan@bbsrc.ac.uk; fax 44-1603-450025.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.3723405.

15:1632–1642 ©2005 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/05; www.genome.org

rently represented by the TIGR (The Institute for Genome Research) v5 analysis (Haas et al. 2005) (http://www.tigr.org/tdb/e2k1/ath1/ath1.shtml). A comprehensive reassembly of the genome incorporated newly sequenced BACs from centromeric regions and carefully examined clone overlaps to yield 119 Mb of analyzed sequence. Genes in this new assembly were then predicted using FL-cDNA and EST sequences, resulting in 19,117 genes matching assemblies of EST and FL-cDNA sequences. About 800 new genes were identified in intergenic space. A total of 26,207 protein-coding genes encode 27,885 distinctive proteins by alternative splicing, nearly all of which contain known protein domains. Approximately 36% of the predicted proteome is encoded by segmental and tandem genomic duplications. Approximately 1400 pseudogenes were identified, many of which had degenerated protein sequences and premature stop codons compared to the family members to which they were most closely related. Improved analysis of repeat and transposon sequences helped to identify 2355 transposon loci. Many of these had been defined as protein-coding genes in the original annotation. Table 1 summarizes some of the analyses of the TIGR v.5 annotation (Haas et al. 2005) and other data.

The new set of predicted genes was functionally annotated using conserved domain composition, not overall sequence homology as performed originally (The *Arabidopsis* Genome Initiative 2000). This permits protein families and relationships between proteins to be more thoroughly defined. Gene Ontology (Ashburner et al. 2000) terms were used to describe each gene in terms of the molecular function of the encoded protein, the biological process in which the protein functions, and the cellular component to which the protein may belong. All proteins were manually assigned to at least one of these categories by these authors (Haas et al. 2005). These manual and computational assignments will continue to be refined as further evidence of gene functions are determined.

An Affymetrix whole-genome array (WGA) for the *Arabidopsis* genome was hybridized with cRNA to detect transcripts and their chromosomal location (Yamada et al. 2003). This landmark analysis supported ~5000 hypothetical gene models as actively transcribed genes, and identified transcripts arising from many intergenic regions and from 20% of the 1300 pseudogenes. Several transcriptional hotspots were identified in centromeric regions that corresponded to a variety of repeats and transposons as well as previously unrecognized genes. Finally, transcription units for most of the genes were accurately defined and used to identify full-length ORF clones for >30% of the genome. These ORFs have been cloned (by the RIKEN/SALK and Agrikola projects) and are critically important resources for functional genomics; their uses are described below.

Table 1. Summary of *Arabidopsis* genome features from current analyses

Genome size	146 Mb (estimated)
Sequenced and annotated genome space	119 Mb
Predicted protein coding genes	26,207
Alternately spliced genes	2330
Protein coding genes with identified transcripts	19,117
Genes in protein families	18,641
Transposons and pseudogenes	3786
Distinct proteins	27,855
Nonredundant cloned ORFs	14,668
Genes with insertions in exon + intron space	24,589

The dynamic genome

One of the major features of the *Arabidopsis* genome revealed by the genome sequence was the extent of gene duplication and segmental duplications, which was surprising given the expectation of a functionally compact genome. Approximately 60% of the genome was thought to be derived from a single duplication event, possibly of the entire genome (The *Arabidopsis* Genome Initiative 2000). Subsequently, more detailed analysis (Ku et al. 2000; Blanc et al. 2003) proposed that the *Arabidopsis* lineage has undergone at least two duplications, the most recent being a polyploidization event during the early evolution of the crucifers (Blanc and Wolfe 2004b). These analyses support a model of *Arabidopsis* genome evolution involving cycles of gene duplication, gene loss, and gene divergence. Genes that remain duplicated tend to become specialized—for example, by different expression patterns (Blanc and Wolfe 2004a). The frequent observation of gene families as tandem arrays further supports the view of a dynamic genome driven by duplication events and specializing gene function from multiple copies of genes. The extensive work carried out based on the *Arabidopsis* genome sequence also supports interpretations of the evolution of the vertebrate lineage that propose a central role for genome duplications (Wolfe 2001). Future work aims to assess and understand genome dynamics, such as gene loss, and altered gene expression that occurs as a consequence of polyploidization. These studies draw on the *Arabidopsis* genome sequence to understand how genome duplications shape evolution and crop plant performance (Osborn et al. 2003).

A. thaliana is a wild species adapted to survive in a wide geographical range, and there is a long legacy of its use as a model for adaptation. Consequently there is an extensive range of natural variation in growth and environmental responsive traits that provide an exceptionally rich source of diversity. Several loci exhibiting variation in complex traits (Quantitative Trait Loci or QTL) have been cloned. Examples include using linkage disequilibrium (LD) to fine map the *FRI* and *FLC* loci controlling flowering time (Hagenblad et al. 2004). Natural variation in hypocotyl responses to light were shown to be due to polymorphisms in phytochrome light receptors (Borevitz and Nordborg 2003). Frequent sequence polymorphisms were identified between the canonical sequenced strain Columbia (Co) and the extensive sequence of another lab strain, Landsberg *erecta* (L*er*) (The *Arabidopsis* Genome Initiative 2000), that are a key resource for map-based cloning. High-throughput methods for identifying polymorphisms have been developed, such as capillary-based SNP detection, to exploit the useful polymorphisms. Affymetrix expression arrays have also been used for genotyping; total genomic DNA from Recombinant Inbred Lines (RILs) made from a cross of Col and L*er* was hybridized to the ATH1 Affymetrix array, and recombination events were identified (Borevitz and Nordborg 2003). Recently, Nordborg et al. (2005) sequenced small genomic regions of 96 accessions of *A. thaliana* to provide a systematic survey of a dense pattern of polymorphisms from a large sample of individuals. This allowed, for the first time, a view of patterns of genome-wide polymorphism and population structure. A clear relationship between geographical distance and genotype distance was shown by clustering allele frequencies. This was quite surprising given that humans have transported *Arabidopsis* around the globe. There is widespread sharing of variation within these populations, and common accessions (ecotypes) share haplotypes though recombination, demonstrating

that these lab strains must not be considered as "lineages." The authors also conclude that the statistical tests commonly used to analyze polymorphism data are not valid in *A. thaliana* because there is "too much local polymorphism." For example, there was a positive correlation between polymorphism levels and genome duplications, and a negative correlation with gene density. These factors require that studies to identify polymorphism frequencies that contribute to phenotypic variation will require surveys of the whole genome. Nevertheless, these authors propose that *A. thaliana* is a good model for evolutionary genomics. Recently, projects were announced by the DOE Joint Genome Institute (JGI) that aim to sequence the genomes of *Arabidopsis lyrata* and *Capsella rubella*, which are close relatives of *A. thaliana*, using a whole-genome shotgun strategy (http://www.jgi.doe.gov/sequencing/cspseqplans2006.html). These sequences will enable significant advances in many aspects of *A. thaliana* biology, for example, by defining ancestral polymorphisms and conserved regions or "footprints" that may have functional significance.

Repeats

One of the reasons *Arabidopsis* was chosen for complete sequencing was its relative lack of repeat sequences compared to other experimentally tractable plants. Cytogenetic analysis (Fransz et al. 2000) revealed extensive tracts of pericentromeric heterochromatin and two rDNA loci at the northern ends of Chromosomes 2 and 4. Chromosome sequencing showed that their pericentromeric regions contained a complex mixture of retroelements, transposons, microsatellites, and middle-repetitive sequence. Unsequenced regions adjacent to and including centromeres contained homopolymeric tracts of characteristic 180-bp and 160-bp repeats. Interstitial heterochromatic regions (or knobs) have been completely sequenced within the context of surrounding low-copy sequences, and these regions have provided deep insights into how heterochromatin is initiated and maintained and how this chromatin state influences gene expression. One of these sequenced regions, hk4S on Chromosome 4, contained a tract of 22.5 tandem copies of the *AtEMSAT1* satellite repeats (Mayer et al. 1999; The Cold Spring Harbor Laboratory, Washington University Genome Sequencing Center, and PE Biosystems *Arabidopsis* Sequencing Consortium 2000). A tiling array was used to profile histone and DNA modifications across these repeats, and it was shown that these and other transposon repeats were methylated and marked by H3mK9 histone modification (Lippman et al. 2004). DNA methylation of these repeats was lost in mutants defective for the *DDM1* chromatin-modifying gene, and this was associated with loss of the H3mK9 mark and increased transcription (Gendrel et al. 2002). The tandem repeats encoded siRNAs, and the levels of these were strongly reduced in the *ddm1* mutant. Furthermore, genes located within the hk4S knob were transcriptionally silent and showed *DDM1*-dependent methylation. It was concluded, based on several well-characterized examples, that epigenetic gene silencing can be mediated by transposable elements inserted in or close to genes. This work was among the first to implicate RNAi in establishing and maintaining epigenetic marks on repeats and genes, and it was directly based on the careful assembly and analysis of complex repeat sequences on *Arabidopsis* Chromosome 4. It has established a new paradigm for understanding how epigenetic marks may be guided to appropriate genome sequences and stably inherited (Martienssen et al. 2005).

Genetic resources

Before the genome sequence was completed, assembled, and annotated, scientists performed chromosome walks to identify mutant loci using YAC and cosmid clones. This approach was slow, uncertain, and had practical drawbacks. The precision and speed with which map-based cloning can now be conducted, given access to the genome sequence and polymorphism data, has greatly accelerated the rate of gene discovery and profitably extended the reach of genetic analysis into many research areas, such as cell biology and metabolism. A plethora of methods and resources have been established over the past five years that provide plant scientists with almost undreamt-of possibilities. Principal among these are the populations of *Arabidopsis* containing T-DNA and transposon insertions. The precise insertion sites of a vast number of these elements in the genome have been determined using PCR-based amplification of flanking sequences and sequencing (Table 2; Alonso et al. 2003; Rosso et al. 2003). These flanking sequences have been aligned with genome sequence assemblies in several databases that permit easy and usually unambiguous identification of genes harboring insertions. To date, there are ~320,000 sequenced insertions in the reference Columbia genome. Lines with sequenced insertion sites are freely available through the *Arabidopsis* Biological Resource Centre (ABRC) and the Nottingham *Arabidopsis* Stock Centre (NASC). These lines are an outstanding resource for functional genomics and should permit the functions of a large number of genes to be determined. This was one of the most important post-genomic objectives of the *Arabidopsis* research community. Future objectives that will further facilitate functional genomics include establishing a bulked set of reference lines with two verified insertion alleles in genes (J. Ecker, pers. comm.). This resource can then be used for systematic surveys of gene function, for establishing populations of crosses to examine genetic interactions, and so on. Other smaller populations containing specialized insertions for the mis-expression of genes, for detecting gene expression patterns as gene or enhancer traps, and for expressing reporter genes from enhancer traps have all been made (Table 2). A significant proportion of the insertions sites have been sequenced, and in most cases the lines have been deposited in the stock centers for distribution.

Many protein-coding genes (~1600) remain with no insertions within exons or introns. Simulations show that doubling the number of random insertions to 600,000 would raise the current proportion of protein-coding genes with insertions from 94% to 98% (Fig. 1). More directed approaches could be taken, for example, using TILLING (Till et al. 2003), RNAi (Lawrence and Pikaard 2003), or newly developed gene replacement strategies (Shaked et al. 2005). This latter method depends on expres-

Table 2. Functional genomics resources in *Arabidopsis thaliana*

Ecotype	Insertion type	Number of sequenced insertions
Wassilewskija	T-DNA	36,287
Landsberg *erecta*	Activation trap	270
Landsberg *erecta*	Enhancer trap	13
Landsberg *erecta*	Gene trap	8891
Landsberg *erecta*	Misexpression trap	899
Columbia	T-DNA	306,168
Columbia	SM	15,943

Data taken from http://www.atidb.org.

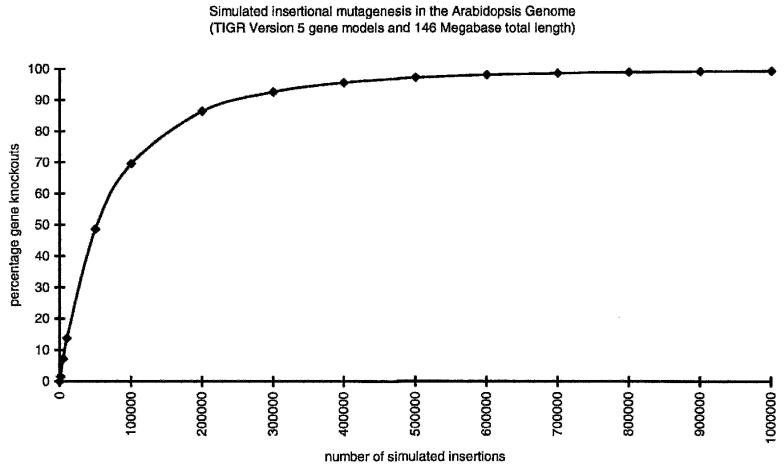

Figure 1. Simulation of random insertions in the *Arabidopsis* genome.

sion of a yeast *RAD54* gene, encoding a member of the *SWI/SNF2* chromatin-remodeling gene family. The gene promotes strand invasion in yeast and is required in that organism for efficient recombination. Expression in *Arabidopsis* increased homologous recombination frequencies by 27-fold, yielding a useful frequency of between 0.01 and 0.1, making it feasible to screen transformants directly by PCR. This method will prove to be very useful and will undoubtedly be further enhanced. Finally, fast neutrons are emerging as the mutagen of choice, as calibrated doses make deletions of corresponding size and frequency (Li et al. 2001). Small deletions targeting loci can be detected in multiplexed samples using PCR, and populations with finely mapped deletions providing coverage of the genome can be made. It is also efficient and feasible to screen deletion lines by hybridization of genomic DNA to Affymetrix chips to map deletions. This strategy has been used in *Medicago truncatula* to identify genes with reduced transcription caused by nonsense codons generated by EMS mutagenesis (Mitra et al. 2004). Taken together, the resources for functional genomics that have been developed and made available provide the entire plant research community with many of the resources needed for progress toward the goal of describing the functions of every gene by 2010.

Gene expression

Array- based transcript detection methods were devised soon after the completion of the genome sequence. Perhaps the most widely used system is the Affymetrix ATH1 array, which has probe sets representing ~24,000 genes as 11 25-mer probe pairs per gene on a single chip. Full-length cDNAs and long gene-specific oligonucleotides have also been printed onto microarrays and used to analyze gene expression. A comprehensive set of genome sequence tags (GSTs) has also been created (Hilson et al. 2004) in the CATMA (Complete *Arabidopsis* Transcriptome Analysis) project (http://www.catma.org/). These are PCR-amplified regions of genomic DNA between 150 and 300 bp long that provide gene-specific probes to ~21,000 *Arabidopsis* genes. These have been used in transcript profiling experiments with printed microarrays (Hilson et al. 2004), and PCR amplicons and cloned GSTs are available from NASC. GSTs were also configured in GATEWAY vectors and used in RNAi experiments to down-regulate gene expression. These reagents provide a key resource for functional genomic experiments.

New insights into the role of RNA species in regulating gene expression and genome dynamics are being gained rapidly based on reference to the genome sequence and annotation. Massively parallel signature sequencing (MPSS) identified a large number of non-coding RNAs (ncRNA) (Meyers et al. 2004), as did whole-genome arrays (Yamada et al. 2003). These data also revealed extensive antisense transcription of genes (Wang et al. 2005) and micro-RNAs, most of which appear to be encoded in non-gene space and may act as regulators of gene expression (Reinhart et al. 2002). They may serve as sequence templates to direct chromatin protein complexes to specific locations, as has been suggested for repeats (Martienssen and Colot 2001; Gendrel et al. 2002; Lippman et al. 2004). One of the exciting challenges in this area, to which an accurately annotated genome sequence can contribute, is establishing the role of small RNA species in growth and development by determining their roles in controlling gene expression through chromatin remodeling and translational control. The machinery producing ncRNAs is starting to be understood—for example, one of the surprises arising from analysis of the genome sequence was an outlier RNA polymerase dubbed RNA Pol IV (The *Arabidopsis* Genome Initiative 2000). This has recently been shown to be a silencing-specific RNA polymerase required for production and maintenance of small interfering RNA (siRNA) (Herr et al. 2005).

Many microarray experiments have been conducted that provide a large amount of quantitative data on gene expression in different *Arabidopsis* tissues and in response to different treatments and experimental conditions. Data from purified cell types (Birnbaum et al. 2003) and microdissected tissues (Casson et al. 2005) continue to extend the resolution of array analysis to the cellular level. Recently, Zanetti et al. (2005) immunopurified polysomes using an epitope-tagged ribosomal protein and showed that transcript profiles obtained by RNA isolated from these polysomes yielded highly reproducible data. Thus it may be possible to express epitope-tagged ribosome proteins from cell-specific promoters (e.g., from enhancer trap lines) to fractionate cell-specific mRNA populations. Finally, it is conceivable that microarray data can be integrated with reporter gene expression patterns, for example, derived from gene trap lines (Sundaresan et al. 1995), to establish a comprehensive and cell-specific analysis of gene expression.

Microarray data from *Arabidopsis* experiments are accumulating rapidly in different databases such as ArrayExpress and NASC. For example, the ATGENEXPRESS network (http://web.uni-frankfurt.de/fb15/botanik/mcb/AFGN/atgenex.htm) recently integrated the results from ~1300 microarray experiments in databases at TAIR and NASC. Data from different experiments can be analyzed because data submitted to these databases are MIAME compliant—that is, there are standard descriptions of experimental conditions, hybridization conditions, and so on that permit statistically sound comparisons. Tools such as GENVESTIGATOR have been developed to analyze sets of *Arabidopsis* microarray data (Zimmermann et al. 2004). This system provides sophisticated data-mining and other analytical tools to identify

expression profiles of sets of genes in different conditions and tissues. These data can also be downloaded for further analyses. So far, array experiments have mostly been interpreted qualitatively—for example, to describe the classes of genes regulated by light, stress, and so on.

The comprehensive and quantitative data obtained from multiple chip experiments can also be used to establish functional modules based on coregulation and tissue specificity and to model gene expression networks. For example, gene expression data from ATH1 chips of genes involved in metabolism have been mapped to corresponding pathways to interpret metabolic consequences of gene expression (Thimm et al. 2004). Recently a larger study (Schmid et al. 2005) integrated data from 79 experiments that assayed RNA levels in different tissues and organs of *Arabidopsis* plants. An expression map of the classes of genes expressed in different organs and tissues was established by reference to GO classifications of gene function. This type of meta-analysis is an initial step in establishing a quantitative framework for interpreting the influence of mutations on gene expression and for defining functional modules. Similar expression maps have been made for yeast (Hughes et al. 2000), *C. elegans* (Kim et al. 2001), and mouse (Zhang et al. 2004). These studies associated coregulated genes and GO descriptions of gene expression to establish functional modules using clustering analysis. Coclustering of genes with known and unknown functions ("guilt by association") can provide important clues to the functions of many genes, which can subsequently be experimentally defined. Such analyses, when coupled to comprehensive GO annotations and interactome data from *Arabidopsis* (see below), will provide important quantitative information about cellular processes. These analyses currently provide an "averaged" view of tissues that are composed of several cell types, and resolution at the cellular level is required to provide key data on the subfunctionalization of cells and how cell functions are integrated to generate tissue and ultimately whole plant responses. This issue is discussed further below in the context of cell and systems biology in *Arabidopsis*.

Array data have also started to be used for determining transcriptional networks, that is, the interplay of transcription factors and promoter sequences that directs the place, time, and level of gene transcription. Establishing models of transcription networks is critically important for defining the logic of regulatory sequences. Experimental approaches in yeast (Lee et al. 2002) used chromatin immunoprecipitation to identify the genome localization of 106 epitope-tagged transcriptional regulators. The promoter regions of ~2300 genes were tagged by at least one regulator, many were tagged by several regulators, and each regulator bound on average 38 promoter regions. Models of regulatory networks such as feed-forward loops and regulatory chains were established based on these data. Currently, it is challenging to establish regulatory networks using this strategy in *Arabidopsis*, as it has ~1700 putative DNA-binding transcription factors and a much larger genome, and multicellularity will complicate interpretation. Nevertheless, the promise of this type of analysis suggests it will be increasingly used in *Arabidopsis* to establish transcriptional regulatory networks.

Computational methods can also generate reasonable models of promoter functions that promise to reduce the complexity of experiments that involve vast numbers of transcription factors such as found in *Arabidopsis*. Algorithms have been developed that correlate promoter motifs with expression patterns. One method involves clustering genes according to their expression levels in multiple conditions and identifying promoters of genes in coregulated clusters. Global alignment, frequency analysis, and other methods have then been used to identify sequence motifs common to each cluster (Tavazoie et al. 1999; Bussemaker et al. 2001). The motifs can then be related to the functions of known promoter motifs (e.g., those in the PLACE database) (Higo et al. 1999) and experimentally tested to determine their function. The application of these strategies in *Arabidopsis* holds great promise in interpreting gene expression data and generating models of gene expression networks. Of particular relevance to this type of analysis is the identification of promoter sequences conserved between closely related species. These phylogenetic footprints may identify functionally relevant promoter domains, for example, comparison of the *CHS* and *AP3* promoter sequences across relatives of *Arabidopsis* identified conserved domains that were functionally significant (Koch et al. 2001). Whole-genome sequencing and assembly of *A. lyrata* and *C. rubella*, together with the planned sequence of *Brassica* species, will provide new opportunities for the systematic determination of gene regulatory architecture in *A. thaliana* as has been proposed for yeast (Kellis et al. 2003).

Determining gene function

The foundations for systematic determination of gene function have been laid by the reannotation of the genome (Haas et al. 2005), the generation of >14,000 nonredundant full-length cDNA clones (Seki et al. 2002) (http://signal.salk.edu/cgi-bin/tdnaexpress; http://rarge.gsc.riken.go.jp/cdna/cdna.pl), the definition of transcription units (Yamada et al. 2003; Meyers et al. 2004), and the development of mathematically structured and systematic descriptions of gene functions (http://www.arabidopsis.org/tools/bulk/go/index). As described above, the initial annotation of the genome was based on limited full-length cDNA sequences. Currently there are 14,668 nonredundant cDNAs available from the RIKEN Bioresource Center (BRC) database made in collaboration between the SSP (Salk, Stanford and Plant Gene Expression Laboratory) and RIKEN groups. Many of these ORFs are cloned into a versatile vector that permits recombination-based manipulation of ORFs into epitope tags, yeast two-hybrid baits, and so on. These collections provide an indispensable resource for large-scale screens, for example, for determining the interactome, systematic determination of biochemical functions, and protein localization. Two large-scale studies have pioneered the systematic determination of biochemical activities of many members of the glycosyl transferase (Bowles et al. 2005) and cytochrome P450 protein families (Schuler and Werck-Reichhart 2003) that catalyze many diverse biochemical reactions in plants. Smaller-scale studies have shown the utility of using ORF-GFP fusions to determine the subcellular localization of proteins (Cutler et al. 2000; Koroleva et al. 2005), and these can be scaled up for cell-based assays using transient expression systems.

Several studies have described in some detail the relationships and functions of genes in large families and have incorporated experimental data from the literature. These include important studies of the MYB (Stracke et al. 2001), bZIP (Jakoby et al. 2002), and b-HLH (Toledo-Ortiz et al. 2003) transcription factor families, the CDPK-SnRK superfamily of protein kinases (Hrabak et al. 2003), pentatricopeptide repeat proteins (Lurin et al. 2004), and receptor-like kinases (Shiu et al. 2004). Many enzymes of primary metabolism have been carefully annotated and

incorporated into specialist databases such as AraCYC (Mueller et al. 2003). Enzymes involved in carbohydrate metabolism have been thoroughly analyzed and annotated (http://afmb.cnrs-mrs.fr/CAZY/). These studies and reviews are critically important for functional genomics studies, as they engage experts in systematic analyses of large numbers of genes. These studies would have even more value if the data were expressed in standardized ontological terms and incorporated into a common framework such as offered by genome databases.

Proteomics strategies are increasingly used in *Arabidopsis* research, and the power of these methods has been greatly improved by the careful reannotation of the genome. Proteins found in the chloroplast, mitochondria, peroxisome, cell wall, nucleus, vacuoles, nucleolus, and cytoskeleton have all been carefully cataloged (Baginsky and Gruissem 2004; Heazlewood and Millar 2005; Pendle et al. 2005) and provide many important clues for future functional studies. Purification and mass-spectrometry of epitope-tagged proteins are increasingly used to identify complexes and post-translational modifications. In other model eukaryotes, notably yeast and *C. elegans*, "interactomes," or networks of protein–protein interactions, have been established by massive yeast two-hybrid screens using ORFs cloned in recombinational systems. Analysis of *C. elegans* interactomes (Li et al. 2004) identified functionally significant interactions that suggested functions for many proteins. Clearly, the availability of a large and well-characterized ORF collection in *Arabidopsis* makes the generation and analysis of *Arabidopsis* interactomes an important and feasible priority. The integration of protein interaction, genetic, and transcriptional networks has revealed global patterns of connections that have generated important biological insights. In yeast (Zhang et al. 2005) networks of interactions, such as protein–protein interactions and gene co-expression data, corresponded with known complexes (replication, actin associated, chromatin remodeling, etc.) and interactions between complexes (actin-associated proteins, motor proteins, protein translocation, etc.). Regulonic complexes linking transcription factors and their regulation of components of complexes were also established. These powerful types of analyses will be possible in *Arabidopsis* once interactome data are generated.

Comparative genomics and crop plant research

Table 3 shows the current state of sequencing in plants, mosses, and green algae. This includes the first planned sequence of a member of the asterid family, tomato (*Lycopersicon esculentum*), which will complement the five rosid species being sequenced. A broader sequence survey of plant taxa, including key species representing important nodes in plant evolution, is also under way (Pryer et al. 2002). This sampling of taxa strongly aids taxonomy and evolutionary studies and is also relevant to identifying plant biodiversity (Wheeler et al. 2004).

Recently the map-based sequence and analysis of the rice genome has been published (The International Rice Genome Sequence Project 2005) and is described elsewhere in this volume (Paterson et al. 2005). Comparison of the rice and *Arabidopsis* proteomes showed that 71% of predicted rice proteins were reasonably similar to *Arabidopsis* proteins. While much more analysis needs to be done to determine relationships, this promising and somewhat unexpected high similarity suggests that the cellular and biochemical functions of many rice genes can be interpreted according to experiments conducted in *Arabidopsis*. Thus, determining orthologous relationships between *Arabidopsis* and crop plant genes is a sound way of translating information on *Arabidopsis* gene function and is therefore a high priority. The same strategy applied to the current annotation of the *Arabidopsis* genome (Haas et al. 2005), based on defining common domain architectures and their conservation, could be applied to sequenced plant genomes as part of a concerted comparative genomics study. The extensive synteny observed between *Arabidopsis*, *M. truncatula*, and soybean (Mudge et al. 2005) suggests that novel bioinformatics approaches can be developed to take into

Table 3. Public plant genome sequencing projects

Group	Organism	EST	Gen.	Comment
Flowering plants				
Asterids	Tomato	+	+	Map-based sequencing initiated
	Tobacco	+		
	Potato	+	+	Map-based sequencing planned
Rosids	*Arabidopsis thaliana*	+	+	Map-based sequencing completed 2000
	Arabidopsis lyrata		+	Whole-genome shotgun starts 2006
	Capsella rubella (Shepherd's Purse)		+	Whole-genome shotgun starts 2006
	Brassica	+	+	Map-based sequencing planned
	Cotton	+		
	Soybean	+		
	Lotus	+	+	Map-based sequencing underway
	Mimulus (Monkey Flower)		+	Whole-genome shotgun starts 2006
	Medicago (barrel medic)	+	+	Map-based sequencing underway
	Cottonwood/aspen	+	+	Whole-genome shotgun completed 2004
Monocots	Barley	+		
	Rice	+	+	Map-based sequencing completed 2000
	Sorghum	+	+	Whole-genome shotgun starts 2006
	Wheat	+		
	Maize	+	+	Map-based sequencing announced
Conifers	Pinus	+		
	Gnetum	+		
Club moss	Selaginella	+	+	Whole-genome shotgun underway
Moss	Physcomitrella	+	+	Whole-genome shotgun underway
Green algae	Chlamydomonas		+	Whole-genome shotgun completed 2005

account the chromosomal context of genes in determining ancestral relationships and relationships between paralogous families.

Several aspects of *Arabidopsis* biology have provided unexpected yet important knowledge about crop plant biology. Its small stature and annual growth habit suggested *Arabidopsis* had little to offer to understanding wood formation in trees. Nevertheless, inflorescence stems undergo secondary thickening and a bona fide cambium forms, and many genes expressed during secondary thickening and associated with cambial activity are highly conserved between *Arabidopsis* and *Populus* (Hertzberg et al. 2001). Furthermore, the systematic characterization of genes involved in cell wall formation, lignification, and vascular cell identity in *Arabidopsis* provides key leads for an important biotechnology sector. Although *Arabidopsis* and other members of the Brassicaceae do not engage in symbiotic interactions, studies in the legume species *Lotus japonicus*, *M. truncatula*, and pea have identified classes of proteins with conserved functions such as those involved in Ca^{2+}-mediated signaling, polar cell growth, and trimeric G-protein-mediated signaling (Limpens and Bisseling 2003). Extensive analysis in *Arabidopsis* of the function of this class of protein in responses to growth regulators and pathogens provides a strong framework for dissecting their role in early nodulation events.

Bioinformatics and modeling

Arabidopsis researchers are supported by the hub provided by TAIR (The *Arabidopsis* Information Resource) and an extensive set of specialist genome databases that distribute *Arabidopsis* data (Rhee et al. 2003). Currently, a dedicated group is annotating *Arabidopsis* and other plant proteins centrally (ftp://ftp.arabidopsis.org/home/tair/Genes/Gene_Ontology/), and these GO terms describing gene functions are starting (albeit quite slowly) to replace the ad hoc system of gene descriptions currently widely used. GO terms describe proteins according to molecular function, biological process, and cellular component (Ashburner et al. 2000). These generic GO classifications provide a strong unifying concept in genome analysis within *Arabidopsis* and other plant species and between all genomes (http://www.godatabase.org/cgi-bin/amigo/go.cgi). Plant Ontology and Trait Ontology have also established controlled vocabularies describing plant anatomy and development, and plant traits and phenotypes (see http://www.gramene.org/plant_ontology/), while Structure Ontology aims to unify the description of genome features allowing for rapid feature transfer to newly sequenced genomes (http://song.sourceforge.net/).

An increasing number of large-scale studies relate biological data to genome features. To manage this work, powerful and useful generic software and database schemas have been developed by the Generic Model Organism Project (GMOD; http://www.gmod.org/). Chief among these is the Genome Browser. It provides a powerful application programming interface that allows sophisticated queries and reports to be constructed with relative ease. The system also provides a facility to allow nonprogrammers to overlay and visualize private data (Stein et al. 2002). Genome Browser has been used for large-scale functional genomics studies (e.g., http://www.atidb.org) and research on epigenetic control of chromosome-scale features (Martienssen et al. 2005). The benefits of data capture and dissemination from distributed sources such as specialist genome databases include direct access to up-to-date data generated by experts, but this creates problems for users who need to know where information can be found, how to relate different data sets, and how to overcome incompatibilities between different systems. The BioMOBY system enables interoperability between database providers through a Web service-like architecture. Services are registered at a central repository that allows bespoke reports to be generated with the visual programming tool Taverna (http://taverna.sourceforge.net/). This method has been successfully implemented by the PlaNet database Consortium (Wilkinson et al. 2005), who provide a comprehensive array of *Arabidopsis* genome-related data resources through BioMOBY services (http://mips.gsf.de/projects/plants/PlaNetPortal/index_html).

One of the major challenges and responsibilities of researchers working on reference organisms such as *Arabidopsis* is the need systematically to capture the large amount of functional genomics data being generated. This permits all plant researchers to make links and integrate diverse types of data to obtain a global perspective of gene functions, as it is now well understood that networks of interactions rather than individual proteins themselves provide a better description of how biological processes are regulated. Currently 24 papers per working day are produced that mention *Arabidopsis*. This vast amount of information is not able to be captured by manual curation strategies alone, thus these data are fragmented, dispersed, and buried in the literature. Furthermore, the data are represented by a myriad of unrelated terms so that they cannot be accessed or manipulated computationally. In response to the problems created by traditional publication methods, text-mining tools are being developed to resurrect data for computational manipulation. However, these approaches are currently limited by restricted access to the full text of journal articles (Krallinger and Valencia 2005). The adoption of semantic tags by learned journals, which enable computers to mine text (Mons 2005), would also be a huge advance in this respect.

This fragmentation occurs at a time when many aspects of *Arabidopsis* research increasingly require individual laboratories to manipulate a common set of data and query all the data at once. The current display of data from several sources such as SIGnAL, TAIR, MIPS, ATIDB, and AtEnsembl further exacerbates the fragmentation problem. A strategy that promotes and rewards a division of labor between databases could help to overcome one aspect of this problem, that is, the collection, processing, analysis, and redistribution of major data sets. This requires free and open access to published community data, including software and backend database systems.

Once data can be successfully captured, they need to be integrated with existing knowledge to generate new hypotheses. Pathways are well understood biologically and computationally in graph theory and provide a strong framework upon which to build this integrated knowledge (Cary et al. 2005). The AraCYC (Mueller et al. 2003) database is a carefully curated database of metabolic pathways that could be expanded to include a more extensive network of metabolic reactions. However, it is difficult for such a system to form the basis for a community effort because its licensing terms are currently inconsistent with the need to share and develop data and databases in an open source environment. The Reactome system (http://www.reactome.org) does provide a framework for describing the molecular biology of an organism (the reactome). The initial implementation is for the human genome (Joshi-Tope et al. 2005) and incorporates a data model based on the concepts of reaction and pathway to describe

molecular processes, such as signal transduction, gene regulation, and metabolism up to higher-level processes. Data are captured by expert curators, and pathways are projected onto other organisms to provide a framework for inferring gene functions. Reactome captures all actors in molecular processes as instances of database entities. This goes closer to modeling biological reality than is possible with systems like Gene Ontology or MapMan (Thimm et al. 2004) in their current invocations, allowing data sets from hybridization arrays and mass-spectroscopy experiments in metabolomics and proteomics to be overlaid on these pathways. In this way high-throughput and highly parallel data sets can be interpreted as correlations with existing knowledge. *Arabidopsis* data are currently being annotated into Reactome, and individuals can also use the curatorial tools to establish pathways in which they are experts, but the degree of evidence for particular data sets requires careful consideration or the adoption of specific evidence ontologies.

One of the most important and challenging aspects of *Arabidopsis* research is to translate knowledge for use in crop plants and to establish evolutionary relationships. Establishing orthologous relationships among the proteomes of the sequenced genomes of crop plants is therefore becoming a high priority. Genome and segmental duplication events during the evolution of flowering plants, and the abundance of large gene families, indicates that defining gene relationships will be as demanding as it will be rewarding. OrthoMCL (Li et al. 2003) establishes similarity matrices using BLASTP within and between species, and then uses Markov clustering to resolve multiple relationships in similarity space. Complete annotations are available for several eukaryotes, including *Arabidopsis* (http://www.cbil.upenn.edu/gene-family). The recent completion of a high-quality rice genome sequence provides the first major opportunity to establish orthologous relationships between rice and *Arabidopsis* proteomes.

The availability of gene catalogs and systematic descriptions of *Arabidopsis* gene function provide plant scientists with the opportunity to develop high-throughput assays to determine cellular readouts from the activities of multiple genes (Gibon et al. 2004). These studies can then be used to develop quantitative models of pathways and other cellular phenotypes to test predictions and develop new experiments. L-Systems have been developed to provide a quantitative framework for describing plant architecture and how this changes during growth and development (Prusinkiewicz 2004), and for describing physiological responses (Allen et al. 2005). This approach describes plants as a set of modules, each of which has a single mathematical description. Variables such as growth rate, genetic regulatory networks, and so on can be incorporated into these modules and the model run to "grow" the virtual plant (Fig. 2). The effects of mutations and gene interactions on growth can then be described according to the growth model. This provides a robust universal framework for describing growth and development phenotypes. At a cellular level, models of gene action in the shoot apical meristem have been established (Jonsson et al. 2005). These incorporate cellular lineage data (Reddy et al. 2004) and information about the spatial and temporal interactions of *CLAVATA1*, *CLAVATA3*, and *WUSCHEL* that maintain expression zones in the shoot apical meristem. This model was able to simulate changes in the *WUSCHEL* domain seen in experiments. These pioneering efforts in quantitative modeling of whole-plant and cell-based growth and development provide foundations for more extensive work aimed at modeling organ development at a cellular level. The planar shape of leaves (Rolland-Lagan et al. 2003) is a promising area in which significant progress is already being made.

Figure 2. L-Systems model of *Arabidopsis* inflorescence growth. This graphic shows a rendering of an L-systems model of the flowering shoot apex of *Arabidopsis*. This figure was generated by Przemek Prusinkiewicz, Enrico Coen, and their colleagues.

Perspectives

Looking forward, the prospects for plant science appear to have never been better. The *Arabidopsis* and rice genomes, and the genome sequences currently being generated and analyzed (Table 3), provide a strong platform for supporting integrative plant science across model and crop species. The complete and accurate sequence of reference genomes from the major groups provides a framework for using sequence from promising high-throughput methods (Margulies et al. 2005) for gene discovery in many new groups of plants. Natural variation can be explored to understand adaptation (Weigel and Nordborg 2005) and broaden the scope of plant breeding (Gur and Zamir 2004). Access to extensive *Arabidopsis* functional genomics resources (Alonso et al. 2003) promotes all plant researchers to consider developing research programs with genetics at their core. New types of screens can be developed in plants using cell-based systems to dissect regulatory pathways and high-throughput RNAi-based methods (DasGupta et al. 2005). Stomatal and trichome cells are good candidates for systematically determining sets of genes involved in signal transduction and cell shape.

These unprecedented opportunities are coupled to socioeconomic trends suggesting a greater need for plant research. For example, more people deserve access to higher-quality food, and plant research can help promote improved plant productivity. Agriculture consumes most of the available high-quality fresh water, and plant research may be able to promote more efficient use of this precious resource. Currently only a small proportion of plant biomass is directly used for fuel and fiber. Increased atmospheric CO_2 levels, dwindling fossil fuel reserves, and their increasing costs suggest that we now need to accelerate research plans to make greater use of plant-based biomass as a renewable chemical feedstock and for energy production.

Acknowledgments

We thank Przemek Prusinkiewicz, Enrico Coen, and colleagues for Figure 2. This work was funded by the Core Strategic Grant to

the John Innes Centre and EC grant QRL1-CT-2001-00006 (PlaNet) to M.B. and S.W.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287:** 2185–2195.

Allen, M.T., Prusinkiewicz, P., and DeJong, T.M. 2005. Using L-systems for modeling source-sink interactions, architecture and physiology of growing trees: The L-PEACH model. *New Phytol.* **166:** 869–880.

Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P., Stevenson, D.K., Zimmerman, J., Barajas, P., Cheuk, R., et al. 2003. Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* **301:** 653–657.

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

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25:** 25–29.

Baginsky, S. and Gruissem, W. 2004. Chloroplast proteomics: Potentials and challenges. *J. Exp. Bot.* **55:** 1213–1220.

Birnbaum, K., Shasha, D.E., Wang, J.Y., Jung, J.W., Lambert, G.M., Galbraith, D.W., and Benfey, P.N. 2003. A gene expression map of the *Arabidopsis* root. *Science* **302:** 1956–1960.

Blanc, G. and Wolfe, K.H. 2004a. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* **16:** 1679–1691.

———. 2004b. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell* **16:** 1667–1678.

Blanc, G., Hokamp, K., and Wolfe, K.H. 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.* **13:** 137–144.

Borevitz, J.O. and Nordborg, M. 2003. The impact of genomics on the study of natural variation in *Arabidopsis*. *Plant Physiol.* **132:** 718–725.

Bowles, D., Isayenkova, J., Lim, E.K., and Poppenberger, B. 2005. Glycosyltransferases: Managers of small molecules. *Curr. Opin. Plant Biol.* **8:** 254–263.

Bussemaker, H.J., Li, H., and Siggia, E.D. 2001. Regulatory element detection using correlation with expression. *Nat. Genet.* **27:** 167–171.

Cary, M.P., Bader, G.D., and Sander, C. 2005. Pathway information for systems biology. *FEBS Lett.* **579:** 1815–1820.

Casson, S., Spencer, M., Walker, K., and Lindsey, K. 2005. Laser capture microdissection for the analysis of gene expression during embryogenesis of *Arabidopsis*. *Plant J.* **42:** 111–123.

The *C. elegans* Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282:** 2012–2046.

The Cold Spring Harbor Laboratory, Washington University Genome Sequencing Center, and PE Biosystems *Arabidopsis* Sequencing Consortium. 2000. The complete sequence of a heterochromatic island from a higher eukaryote. *Cell* **100:** 377–386.

Cutler, S.R., Ehrhardt, D.W., Griffitts, J.S., and Somerville, C.R. 2000. Random GFP:cDNA fusions enable visualization of subcellular structures in cells of *Arabidopsis* at a high frequency. *Proc. Natl. Acad. Sci.* **97:** 3718–3723.

DasGupta, R., Kaykas, A., Moon, R.T., and Perrimon, N. 2005. Functional genomic analysis of the Wnt-wingless signaling pathway. *Science* **308:** 826–833.

Fransz, P.F., Armstrong, S., de Jong, J.H., Parnell, L.D., van Drunen, C., Dean, C., Zabel, P., Bisseling, T., and Jones, G.H. 2000. Integrated cytogenetic map of chromosome arm 4S of *A. thaliana*: Structural organization of heterochromatic knob and centromere region. *Cell* **100:** 367–376.

Gendrel, A.V., Lippman, Z., Yordan, C., Colot, V., and Martienssen, R.A. 2002. Dependence of heterochromatic histone H3 methylation patterns on the *Arabidopsis* gene DDM1. *Science* **297:** 1871–1873.

Gibon, Y., Blaesing, O.E., Hannemann, J., Carillo, P., Hohne, M., Hendriks, J.H., Palacios, N., Cross, J., Selbig, J., and Stitt, M. 2004. A robot-based platform to measure multiple enzyme activities in *Arabidopsis* using a set of cycling assays: Comparison of changes of enzyme activities and transcript levels during diurnal cycles and in prolonged darkness. *Plant Cell* **16:** 3304–3325.

Gur, A. and Zamir, D. 2004. Unused natural variation can lift yield barriers in plant breeding. *PLoS Biol.* **2:** e245.

Haas, B.J., Wortman, J.R., Ronning, C.M., Hannick, L.I., Smith Jr., R.K., Maiti, R., Chan, A.P., Yu, C., Farzad, M., Wu, D., et al. 2005. Complete reannotation of the *Arabidopsis* genome: Methods, tools, protocols and the final release. *BMC Biol.* **3:** 7.

Hagenblad, J., Tang, C., Molitor, J., Werner, J., Zhao, K., Zheng, H., Marjoram, P., Weigel, D., and Nordborg, M. 2004. Haplotype structure and phenotypic associations in the chromosomal regions surrounding two *Arabidopsis thaliana* flowering time loci. *Genetics* **168:** 1627–1638.

Hall, A.E., Fiebig, A., and Preuss, D. 2002. Beyond the *Arabidopsis* genome: Opportunities for comparative genomics. *Plant Physiol.* **129:** 1439–1447.

Heazlewood, J.L. and Millar, A.H. 2005. AMPDB: The *Arabidopsis* mitochondrial protein database. *Nucleic Acids Res.* **33:** D605–D610.

Herr, A.J., Jensen, M.B., Dalmay, T., and Baulcombe, D.C. 2005. RNA polymerase IV directs silencing of endogenous DNA. *Science* **308:** 118–120.

Hertzberg, M., Aspeborg, H., Schrader, J., Andersson, A., Erlandsson, R., Blomqvist, K., Bhalerao, R., Uhlen, M., Teeri, T.T., Lundeberg, J., et al. 2001. A transcriptional roadmap to wood formation. *Proc. Natl. Acad. Sci.* **98:** 14732–14737.

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.

Hilson, P., Allemeersch, J., Altmann, T., Aubourg, S., Avon, A., Beynon, J., Bhalerao, R.P., Bitton, F., Caboche, M., Cannoot, B., et al. 2004. Versatile gene-specific sequence tags for *Arabidopsis* functional genomics: Transcript profiling and reverse genetics applications. *Genome Res.* **14:** 2176–2189.

Hosouchi, T., Kumekawa, N., Tsuruoka, H., and Kotani, H. 2002. Physical map-based sizes of the centromeric regions of *Arabidopsis thaliana* chromosomes 1,2, and 3. *DNA Res.* **9:** 117–121.

Hrabak, E.M., Chan, C.W., Gribskov, M., Harper, J.F., Choi, J.H., Halford, N., Kudla, J., Luan, S., Nimmo, H.G., Sussman, M.R., et al. 2003. The *Arabidopsis* CDPK-SnRK superfamily of protein kinases. *Plant Physiol.* **132:** 666–680.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., et al. 2000. Functional discovery via a compendium of expression profiles. *Cell* **102:** 109–126.

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

Jakoby, M., Weisshaar, B., Droge-Laser, W., Vicente-Carbajosa, J., Tiedemann, J., Kroj, T., and Parcy, F. 2002. bZIP transcription factors in *Arabidopsis*. *Trends Plant Sci.* **7:** 106–111.

Jonsson, H., Heisler, M., Reddy, G.V., Agrawal, V., Gor, V., Shapiro, B.E., Mjolsness, E., and Meyerowitz, E.M. 2005. Modeling the organization of the WUSCHEL expression domain in the shoot apical meristem. *Bioinformatics* **21 Suppl 1:** i232–i240.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., et al. 2005. Reactome: A knowledgebase of biological pathways. *Nucleic Acids Res.* **33:** D428–D432.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.

Kim, S.K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J.M., Eizinger, A., Wylie, B.N., and Davidson, G.S. 2001. A gene expression map for *Caenorhabditis elegans*. *Science* **293:** 2087–2092.

Koch, M.A., Weisshaar, B., Kroymann, J., Haubold, B., and Mitchell-Olds, T. 2001. Comparative genomics and regulatory evolution: Conservation and function of the Chs and Apetala3 promoters. *Mol. Biol. Evol.* **18:** 1882–1891.

Koroleva, O.A., Tomlinson, M.L., Leader, D., Shaw, P., and Doonan, J.H. 2005. High-throughput protein localization in *Arabidopsis* using *Agrobacterium*-mediated transient expression of GFP-ORF fusions. *Plant J.* **41:** 162–174.

Krallinger, M. and Valencia, A. 2005. Text-mining and information-retrieval services for molecular biology. *Genome Biol.* **6:** 224.

Ku, H.M., Vision, T., Liu, J., and Tanksley, S.D. 2000. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: Large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl. Acad. Sci.* **97:** 9121–9126.

Lawrence, R.J. and Pikaard, C.S. 2003. Transgene-induced RNA interference: A strategy for overcoming gene redundancy in polyploids to generate loss-of-function mutations. *Plant J.* **36:** 114–121.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., et

al. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298:** 799–804.
Li, X., Song, Y., Century, K., Straight, S., Ronald, P., Dong, X., Lassner, M., and Zhang, Y. 2001. A fast neutron deletion mutagenesis-based reverse genetics system for plants. *Plant J.* **27:** 235–242.
Li, L., Stoeckert Jr., C.J., and Roos, D.S. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13:** 2178–2189.
Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., et al. 2004. A map of the interactome network of the metazoan *C. elegans*. *Science* **303:** 540–543.
Limpens, E. and Bisseling, T. 2003. Signaling in symbiosis. *Curr. Opin. Plant Biol.* **6:** 343–350.
Lippman, Z., Gendrel, A.V., Black, M., Vaughn, M.W., Dedhia, N., McCombie, W.R., Lavine, K., Mittal, V., May, B., Kasschau, K.D., et al. 2004. Role of transposable elements in heterochromatin and epigenetic control. *Nature* **430:** 471–476.
Lurin, C., Andres, C., Aubourg, S., Bellaoui, M., Bitton, F., Bruyere, C., Caboche, M., Debast, C., Gualberto, J., Hoffmann, B., et al. 2004. Genome-wide analysis of *Arabidopsis* pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. *Plant Cell* **16:** 2089–2103.
Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380.
Martienssen, R.A. and Colot, V. 2001. DNA methylation and epigenetic inheritance in plants and filamentous fungi. *Science* **293:** 1070–1074.
Martienssen, R.A., Doerge, R.W., and Colot, V. 2005. Epigenomic mapping in *Arabidopsis* using tiling microarrays. *Chromosome Res.* **13:** 299–308.
Mayer, K., Schuller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Dusterhoft, A., Stiekema, W., Entian, K.D., Terryn, N., et al. 1999. Sequence and analysis of Chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* **402:** 769–777.
Meyers, B.C., Vu, T.H., Tej, S.S., Ghazal, H., Matvienko, M., Agrawal, V., Ning, J., and Haudenschild, C.D. 2004. Analysis of the transcriptional complexity of *Arabidopsis thaliana* by massively parallel signature sequencing. *Nat. Biotechnol.* **22:** 1006–1011.
Mitra, R.M., Gleason, C.A., Edwards, A., Hadfield, J., Downie, J.A., Oldroyd, G.E., and Long, S.R. 2004. A Ca^{2+}/calmodulin-dependent protein kinase required for symbiotic nodule development: Gene identification by transcript-based cloning. *Proc. Natl. Acad. Sci.* **101:** 4701–4705.
Mons, B. 2005. Which gene do you mean? *BMC Bioinformatics* **6:** 142.
Mudge, J., Cannon, S.B., Kalo, P., Oldroyd, G.E., Roe, B.A., Town, C.D., and Young, N.D. 2005. Highly syntenic regions in the genomes of soybean, *Medicago truncatula*, and *Arabidopsis thaliana*. *BMC Plant Biol.* **5:** 15.
Mueller, L.A., Zhang, P., and Rhee, S.Y. 2003. AraCyc: A biochemical pathway database for *Arabidopsis*. *Plant Physiol.* **132:** 453–460.
Nordborg, M., Hu, T.T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., et al. 2005. The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol.* **3:** e196.
Osborn, T.C., Pires, J.C., Birchler, J.A., Auger, D.L., Chen, Z.J., Lee, H.S., Comai, L., Madlung, A., Doerge, R.W., Colot, V., et al. 2003. Understanding mechanisms of novel gene expression in polyploids. *Trends Genet.* **19:** 141–147.
Paterson, A.H., Freeling, M., and Sasaki, T. 2005. Grains of knowledge: Genomics of modern cereals. *Genome Res.* (this issue).
Pendle, A.F., Clark, G.P., Boon, R., Lewandowska, D., Lam, Y.W., Andersen, J., Mann, M., Lamond, A.I., Brown, J.W., and Shaw, P.J. 2005. Proteomic analysis of the *Arabidopsis* nucleolus suggests novel nucleolar functions. *Mol. Biol. Cell* **16:** 260–269.
Prusinkiewicz, P. 2004. Modeling plant growth and development. *Curr. Opin. Plant Biol.* **7:** 79–83.
Pryer, K.M., Schneider, H., Zimmer, E.A., and Banks, J. 2002. Deciding among green plants for whole genome studies. *Trends Plant Sci.* **7:** 550–554.
Reddy, G.V., Heisler, M.G., Ehrhardt, D.W., and Meyerowitz, E.M. 2004. Real-time lineage analysis reveals oriented cell divisions associated with morphogenesis at the shoot apex of *Arabidopsis thaliana*. *Development* **131:** 4225–4237.
Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P. 2002. MicroRNAs in plants. *Genes & Dev.* **16:** 1616–1626.
Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., et al. 2003. The *Arabidopsis* Information Resource (TAIR): A model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.* **31:** 224–228.
Riano-Pachon, D.M., Dreyer, I., and Mueller-Roeber, B. 2005. Orphan transcripts in *Arabidopsis thaliana*: Identification of several hundred previously unrecognized genes. *Plant J.* **43:** 205–212.
Rolland-Lagan, A.G., Bangham, J.A., and Coen, E. 2003. Growth dynamics underlying petal shape and asymmetry. *Nature* **422:** 161–163.
Rosso, M.G., Li, Y., Strizhov, N., Reiss, B., Dekker, K., and Weisshaar, B. 2003. An *Arabidopsis thaliana* T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. *Plant Mol. Biol.* **53:** 247–259.
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.
Schuler, M.A. and Werck-Reichhart, D. 2003. Functional genomics of P450s. *Annu. Rev. Plant Biol.* **54:** 629–667.
Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y., et al. 2002. Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science* **296:** 141–145.
Shaked, H., Melamed-Bessudo, C., and Levy, A.A. 2005. High-frequency gene targeting in *Arabidopsis* plants expressing the yeast RAD54 gene. *Proc. Natl. Acad. Sci.* **102:** 12265–12269.
Shiu, S.H., Karlowski, W.M., Pan, R., Tzeng, Y.H., Mayer, K.F., and Li, W.H. 2004. Comparative analysis of the receptor-like kinase family in *Arabidopsis* and rice. *Plant Cell* **16:** 1220–1234.
Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., et al. 2002. The generic genome browser: A building block for a model organism system database. *Genome Res.* **12:** 1599–1610.
Stracke, R., Werber, M., and Weisshaar, B. 2001. The R2R3-MYB gene family in *Arabidopsis thaliana*. *Curr. Opin. Plant Biol.* **4:** 447–456.
Sundaresan, V., Springer, P., Volpe, T., Haward, S., Jones, J.D., Dean, C., Ma, H., and Martienssen, R. 1995. Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. *Genes & Dev.* **9:** 1797–1810.
Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., and Church, G.M. 1999. Systematic determination of genetic network architecture. *Nat. Genet.* **22:** 281–285.
Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, 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.
Till, B.J., Reynolds, S.H., Greene, E.A., Codomo, C.A., Enns, L.C., Johnson, J.E., Burtner, C., Odden, A.R., Young, K., Taylor, N.E., et al. 2003. Large-scale discovery of induced point mutations with high-throughput TILLING. *Genome Res.* **13:** 524–530.
Toledo-Ortiz, G., Huq, E., and Quail, P.H. 2003. The *Arabidopsis* basic/helix–loop–helix transcription factor family. *Plant Cell* **15:** 1749–1770.
Wang, X.J., Gaasterland, T., and Chua, N.H. 2005. Genome-wide prediction and identification of *cis*-natural antisense transcripts in *Arabidopsis thaliana*. *Genome Biol.* **6:** R30.
Weigel, D. and Nordborg, M. 2005. Natural variation in *Arabidopsis*. How do we find the causal genes? *Plant Physiol.* **138:** 567–568.
Wheeler, Q.D., Raven, P.H., and Wilson, E.O. 2004. Taxonomy: Impediment or expedient? *Science* **303:** 285.
Wilkinson, M., Schoof, H., Ernst, R., and Haase, D. 2005. BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. The PlaNet exemplar case. *Plant Physiol.* **138:** 5–17.
Wolfe, K.H. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat. Rev. Genet.* **2:** 333–341.
Wortman, J.R., Haas, B.J., Hannick, L.I., Smith Jr., R.K., Maiti, R., Ronning, C.M., Chan, A.P., Yu, C., Ayele, M., Whitelaw, C.A., et al. 2003. Annotation of the *Arabidopsis* genome. *Plant Physiol.* **132:** 461–468.
Xiao, Y.L., Malik, M., Whitelaw, C.A., and Town, C.D. 2002. Cloning and sequencing of cDNAs for hypothetical genes from Chromosome 2 of *Arabidopsis*. *Plant Physiol.* **130:** 2118–2128.
Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J., Southwick, A.M., Wu, H.C., Kim, C., Nguyen, M., et al. 2003. Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302:** 842–846.
Zanetti, M.E., Chang, I.F., Gong, F., Galbraith, D.W., and Bailey-Serres, J. 2005. Immunopurification of polyribosomal complexes of *Arabidopsis* for global analysis of gene expression. *Plant Physiol.* **138:** 624–635.
Zhang, W., Morris, Q.D., Chang, R., Shai, O., Bakowski, M.A.,

Mitsakakis, N., Mohammad, N., Robinson, M.D., Zirngibl, R., Somogyi, E., et al. 2004. The functional landscape of mouse gene expression. *J. Biol.* **3:** 21.
Zhang, L.V., King, O.D., Wong, S.L., Goldberg, D.S., Tong, A.H., Lesage, G., Andrews, B., Bussey, H., Boone, C., and Roth, F.P. 2005. Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network. *J. Biol.* **4:** 6.
Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., and Gruissem, W. 2004. GENEVESTIGATOR. *Arabidopsis* microarray database and analysis toolbox. *Plant Physiol.* **136:** 2621–2632.

Web site references

ftp://ftp.arabidopsis.org/home/tair/Genes/Gene_Ontology/; TAIR Gene Ontology.
http://afmb.cnrs-mrs.fr/CAZY/; Carbohydrate Active Enzymes.
http://mips.gsf.de/projects/plants/PlaNetPortal/index_html; PlaNet Consortium.
http://rarge.gsc.riken.go.jp/cdna/cdna.pl; RIKEN.
http://signal.salk.edu/cgi-bin/tdnaexpress; Signal.
http://web.uni-frankfurt.de/fb15/botanik/mcb/AFGN/atgenex.htm; ATGENEXPRESS.
http://www.atidb.org; *Arabidopsis* Insertion Database.
http://www.catma.org/; CATMA.
http://www.cbil.upenn.edu/gene-fami ; OrthoMCL.
http://www.gramene.org/plant_ontology/; GRAMENE.
http://www.jgi.doe.gov/sequencing/cspseqplans2006.html; DOE Joint Genome Institute.
http://www.reactome.org; Reactome.
http://www.tigr.org/tdb/e2k1/ath1/ath1.shtml; The Institute for Genome Research OrthoMCL.
http://www.godatabase.org/cgi-bin/amigo/go.cgi; Gene Ontologies.
http://song.sourceforge.net/; Sourceforge software.
http://www.gmod.org/; GMOD genome browser.
http://taverna.sourceforge.net/; Taverna visual programming tool.