Genomics and plant cells: application of genomics strategies to *Arabidopsis* cell biology

Michael Bevan

*Cell and Developmental Biology Department, John Innes Centre, Colney Lane, Norwich NR4 7UJ, UK
(michael.bevan@bbsrc.ac.uk)*

In this review I seek to describe how the complete catalogue of plant genes and proteins, revealed by genome sequencing, can provide novel insights into cell biology. Many new analytical methods have been developed to digest the flood of genome sequence data, including analysis of the transcriptome, proteome and metabolites. High-throughput analysis of protein targeting and other methods will ascribe new information to proteins and create important links with other large datasets. To fulfil the potential revealed by this genomic information, many challenges have to be met. Among these are organizational changes needed to create common datasets accessible to all scientists, and bioinformatics solutions to capture and integrate diverse datasets. Once harnessed, these new strategies will irrevocably change the way we conduct plant science.

Keywords: *Arabidopsis*; genomics; functional genomics; cell biology

1. CELL BIOLOGY AND GENOME SEQUENCES

For some time the irreducible unit of cell biological investigation has been the cell. The catalogue of all genes and predicted proteins generated by whole-genome sequencing now challenges biologists to redefine the concept of the cell in terms of all its components and their interactions. Genome sequencing has provided the foundation for systematic analyses of gene expression and protein interactions, and for the spatial location of proteins and nucleic acids. Together with catalogues of metabolites and models of regulatory networks, it is possible to imagine appropriate visualization and other computational tools that could emulate cellular processes. But even with lists of all components and molecular interactions it is very difficult to imagine how this information could be synthesized into the emergent properties of a cell. Therefore, the challenge to molecular biologists must be to define the functions of genes and proteins in a cellular and whole-organism context, while cell biologists face the challenge of defining cellular structures and functions using the component parts and their interactions. The benefits of establishing this continuum of knowledge of cellular processes, from individual molecules to whole cells, are clear. Using this strategy, a holistic view of a cell as a self-replicating entity is preserved by studying the functions of all genes and proteins within a cell, while the extensive and precise nature of genome-based information permits the isolation and characterization of components. Genome-based information also provides direct comparison of related and divergent cellular functions in different sequenced organisms, adding depth to our understanding of cellular functions.

In this review I will describe some of the salient differences between plant and animal cells, to aid comparison, and then assess the progress in describing the components of plant cells revealed by genome sequencing. Finally, progress in defining the functions of genes and proteins in cellular and multi-cellular contexts will be reviewed, and problems and bottlenecks will be discussed.

2. PLANT CELLS ARE DIFFERENT

Plant cells are, in many ways, fundamentally different units from those of other eukaryotes. These differences may reflect the independent evolution of solutions to meet the requirements of multicellularity (Meyerowitz 1999), while many are clearly adaptations to the distinctive lifestyle of land plants. The most obvious difference between plant and animal cells is the polysaccharide-rich extracellular matrix of plant cells that constrains cell shape and limits motility. Cell division plays a major role in defining cell and organ shape in plants by defining the plane of cytokinesis. Growth and differentiation of plant cells requires reorganization of the cytoskeleton to direct membrane trafficking to zones of surface growth, and to direct formation of the cell plate during cytokinesis. Cytoskeletal components are generally common between plants and animals, except for the absence in plants of proteins such as vinculin, vitronectin, talin and spectrin that link the cytoskeleton across the plasma membrane to the extracellular matrix (The *Arabidopsis* Genome Initiative 2000). This is consistent with the polysaccharide extracellular matrix of plants, but there may be counterparts to at least some of these proteins. For example, cellulose synthase complexes are thought to reside in the plasma membrane and be directed by the cytoskeleton to weave microfibrils in particular orientations during growth (Delmer 1999). The proteins linking this and other types of plasma-membrane and cell proteins to the cytoskeleton are not

One contribution of 12 to a Discussion Meeting Issue 'The plant cell: between genome and plant'.

yet known. The motive force for plant growth derives from turgor pressure generated by a large number of ion and water pumps in plants. Nearly half of the channel proteins in *Arabidopsis* are aquaporins (water channels), and *Arabidopsis* (and probably most other plants) contains 10-fold more major intrinsic protein water channels than non-plant organisms (The *Arabidopsis* Genome Initiative 2000), reflecting the major role of water in many aspects of plant function. Thick cellulosic cell walls prevent direct contact between plant cells, suggesting novel mechanisms of intercellular communication. One of these is the recruitment of receptor Ser/Thr kinases for putative signalling functions in plants, and no plant homologues of the receptor tyrosine kinases commonly found in animals are found in *Arabidopsis* (The *Arabidopsis* Genome Initiative 2000). Defining the functions of extracellular receptors and their ligands and other potential regulatory proteins within the cell wall, such as the wall-associated kinases (Wagner & Kohorn 2001), will reveal novel features of intercellular communication. One striking example is the apparently continuous cytoplasm of the symplast linked via plasmodesmata, through which can flow regulatory proteins and mRNA (Hake 2001). Other control systems, such as the spatial and temporal regulation of mRNA and protein levels, appear to have much in common with those in other eukaryotes, and only differ in the deployment of different proteins to perform similar functions in plants and metazoans. The definition of gene function in the context of plant cells will not only add necessary dimensions to plant cell biology but will also provide important insights into the evolution of eukaryotic cellular processes.

3. NAMING THE PARTS

The accuracy and completeness of the genome sequence record is of fundamental importance for all subsequent analyses. How good is the record? The genome sequences of yeast, *Caenorhabditis elegans* and *Arabidopsis* were derived by shotgun sequencing of multiply overlapping large-insert clones, directed finishing of gaps and resolution of sequence and assembly ambiguities. In general, two sequence assembly programs, PHRAP and XGAP, were used in these organisms, and the assembled sequence was validated by comparison and iteration. This led to sequence errors of less than 1 in 10 000 bases, sufficient for exacting procedures such as whole-genome SNP detection (The *Arabidopsis* Genome Initiative 2000). Centromeric regions in *Arabidopsis* have not yet been completely sequenced due to difficulties in assembling both tiling paths of BACs and shotgun sequence of BACs. The *Drosophila* genome was sequenced using a whole-genome strategy to 12.8 times coverage and assembled using a scaffold of end-sequenced large-insert clones of different sizes and other landmarks (Adams *et al.* 2000). This yielded a high-accuracy sequence anchored in the genome and accessed repetitive regions. The first draft sequence of the human genome was performed using both whole-genome shotgun and clone-by-clone strategies, and these datasets have been integrated (International Human Genome Sequencing Consortium 2001; Venter *et al.* 2001). The clone-based strategy was published at an average of 7.5-fold average sequence coverage, the whole-genome strategy shotgun sequence yielded fivefold coverage and

clone coverage was 38-fold, and the combined public-derived faux shotgun data and the whole-genome coverage yielded a combined eightfold coverage used in whole-genome assembly. This yielded 73% of the sequence in contigs of more than 30 kb. The present state of human sequence is highly accurate but contains many small gaps and is not yet a complete representation of all human genes. Thus detailed comparison of the genomes of highly different sequenced eukaryotes is possible.

Nearly all genes in multicellular eukaryotes are split by exons, and in the absence of experimental data such as full-length cDNA sequences, protein sequences are derived by probabilistic modelling algorithms such as GENEMARK (Lukashin & Borodovsky 1998). Similar approaches have been used in all sequenced eukaryotes. These predictions, incorporating splice-site consensus sequences, codon usage and other features of known *Arabidopsis* genes, were used to predict protein-coding sequences (Pavy *et al.* 1999). A combination of optimal methods provided accurate predictions for *ca.* 40% of genes, with incorrectly identified initiation and termination exons as common faults. Recently (O. White, personal communication), combinations of optimal gene finding algorithms and the sequence of 5000 full-length cDNA have yielded *ca.* 70% correct gene structures.

Systematic analysis of the protein sequences predicted from the *Arabidopsis* genome sequence have been analysed using automated methods such as PEDANT (Frishman *et al.* 2001), which scores significant homologies, known domains and predicted structures to relate proteins to those of known function. These analyses have revealed many novel protein sequences in plants. About 30% of the genes in *Arabidopsis thaliana* have no significant similarity at the protein level to genes found in any other organism, and a significant proportion of these proteins can be expected to contribute to the plant-specific processes described above. Plant cells also may lack many proteins found in nearly all animals. The extent to which these differences are due to extreme sequence divergence between related genes is not yet clear, but exciting progress in defining remote homologies (described below) may help define proteins that are truly plant-specific and those not found in plants.

About 70% of the predicted proteins encoded by the 25 500 *Arabidopsis* genes have significant homology, defined by BLASTP alignments (The *Arabidopsis* Genome Initiative 2000) to proteins in other organisms, and functional assignments are being made based on these and other similarity matches. How these matches are categorized is described below, but first, more informative methods of defining sequence similarity and assigning putative functions are described.

New *Arabidopsis* gene families are being established on the basis of domain analysis, and many new plant-specific peptide domains are enriching domain databases. Proteins are composed of assemblies of conserved domains and signatures that often comprise functional and structural units, interspersed with variable peptide sequences. The significance of homologies to known signatures can be ascertained more easily than homologies across entire proteins because the functional and structural significance of homologies is defined by similarity to type members of the group. INTERPRO (<http://www.ebi.ac.uk/interpro>) unifies

functional and evolutionary knowledge of domains and sites from several sources such as PROSITE and PFAM. The domains are built up by careful multiple alignments of peptides from many sequenced organisms and are manually curated. Whole-genome analysis of predicted *Arabidopsis* peptides revealed interesting differences in the frequencies of domains; kinase catalytic domains are the most abundant shared motifs, while the immunoglobulin domain was confined to metazoans and the AP2 domain DNA-binding repeat is plant specific (see <http://www.ebi.ac.uk/interpro>). SMART reveals the linear arrangement of domains within a protein and helps define new protein architectures (<http://smart.embl-heidelberg.de/>). Analysis of eukaryotic proteomes using SMART shows that vertebrates have significantly more protein architectures than insects, nematodes or plants, and reveals domain accretion to be a major mechanism of protein evolution. These studies will be important for establishing ancestral relationships between eukaryotic proteins. In plants, a clear example of this is evident from the accretion of chloroplast-targeting domains into a variety of proteins, including those of putative cyanobacterial origin.

The prediction of protein targeting domains for subcellular locations is particularly relevant to cell biological research in plants. Computational methods have identified 6000 nuclear encoded genes that are post-translationally targeted to chloroplasts and mitochondria. The PREDOTAR method distinguishes between mitochondrial and chloroplast targeting peptides and showed that *ca.* 3000 nuclear genes are targeted to each organelle (see <http://www.inra.fr/internet/produits/predotar/>). Other protein domains, for example the 7-TM associated with membrane proteins, are reliable indicators of membrane localization. Software such as CBS can locate putative membrane-associated peptide domains to aid prediction of subcellular localization.

Once protein structures have been defined and relationships between proteins established, proteins with related functions can then be categorized according to their predicted cellular functions. In *Arabidopsis* these categories were based on modifications of the *Escherichia coli* and yeast functional catalogues (Riley 1993; see <http://mips.gsf.de/>). Many classes of *Arabidopsis* genes can be categorized according to this scheme. For example, the enzyme activities of most aspects of primary and secondary metabolism are satisfactorily described and gene products can generally be unambiguously allocated. New categories, such as N assimilation, alkaloid synthesis and other reasonably well defined enzymatic functions in plants have also been made (see <http://mips.gsf.de/>). In so far as a protein's functions can be fully captured by describing its enzymatic or other molecular function, this functional catalogue is accurate and satisfactory, and above all, unifying.

However, the extent to which these matches reflect functional relatedness between organisms as different as man and plant cannot be predicted from sequence similarity alone. The processes of comparison can propagate annotation errors and establish very tenuous and occasionally meaningless associations between genes based on conserved amino acid sequences. Furthermore, the relevance of several of the *E. coli* and yeast categories to protein functions in *Arabidopsis* needs careful examin-

ation, for example the relegation of all cell-wall polysaccharide biochemistry to 'other carbohydrate metabolism activities' does not represent the complexity and importance of these enzymatic functions in plant growth, and the pheromone and osmotic response signal transduction categories in plants are clearly either irrelevant or as yet unidentified (see <http://mips.gsf.de/>).

The functions of most genes can be defined in several different ways with respect to enzymatic activity, transport functions, class of transcription factor, etc. Thus kinase is a precise description of an enzymatic function but is not especially helpful in describing cellular functions in signal transduction and stress responses. In many cases cellular functions cannot be meaningfully defined according to the function of a single protein, but depend upon interactions between proteins and other components. Consequently, biological functions of individual proteins also need to be established in terms of their contributions to functional modules (Hartwell *et al.* 1999). Defining the function and components of these modules will be important in bridging the gap between knowledge of the functions of individual genes and cellular-scale processes, such as cell division and cell wall synthesis.

Two important steps are being made to extend and replace the present first-pass functional catalogues. First, experts in particular gene families are providing more detailed and durable analyses of *Arabidopsis* gene products (see <http://arabidopsis.org/>). These include complex families such as those encoding the plethora of cytochrome P450-like enzymes, receptor kinases and chromatin modification proteins. The growing evidence for extensive conservation of genes and proteins among diverse organisms (e.g. approximately half of the 289 genes with established roles in human disease have high matches in the *Arabidopsis* genome; The *Arabidopsis* Genome Initiative 2000) challenges scientists to develop a systematic description of gene function applicable to a wide variety of organisms. The GO Consortium (<http://www.geneontology.org/>) is producing defined and structured controlled vocabularies suitable for describing gene function in multiple organisms. *Drosophila*, mouse, yeast and *Arabidopsis* are the model organisms currently included in the GO Consortium. The *Drosophila* genome was annotated into GO categories using automated methods, indicating significant potential for application. The goal of the Consortium is to establish dynamic ontologies that constantly take account of new knowledge of protein function. Three categories reflecting different aspects of protein function form the foundation of GO: (i) the biochemical function of proteins (ligand binding, enzyme activity, etc.); (ii) the biological processes in which the protein functions (translation, ATP synthesis, etc.—sometimes roughly equivalent to functional modules); and (iii) the cellular component or place in the cell where a gene functions (Golgi apparatus, vacuolar membrane, etc.). Of course, some proteins function in several processes and cellular structures, and the ontologies reflect these multiple relationships. Currently only idealized cells are used, and specific cell types, organs and whole organisms are not represented. Inputs from systematic analysis of gene function and gene expression patterns (see below) will broaden the scope of describing and relating the functions of genes by using common interconnected descriptions. Rep-

resenting environmental effects on gene expression and gene function, and the role of genes in modulating responses to the environment, is especially relevant to plant-gene description. The standard descriptions of experimental conditions implemented for microarray analysis should be helpful in expanding this class of gene description (<ftp://tairpub:tairpub@ftp.arabidopsis.org/home/tair/microarrays/>).

4. PUTTING THE PARTS TOGETHER

Since the completion of the *Arabidopsis* genome sequence and the first-pass analysis and prediction of gene function, two approaches to defining *Arabidopsis* gene function have emerged. The first is experimental and seeks to understand the functions of a relatively small number of genes in detail using screens for informative phenotypes. This type of work generally uses chemical mutagenesis and gene mapping and, increasingly, RNA interference and insertion mutagenesis as a starting point. Presently insertional mutagenesis strategies will have the capacity for large-scale implementation in *Arabidopsis* and yield a variety of effects on genes that can be detected in high-throughput phenotype screens (Kempin *et al.* 1997). Both of these approaches have much in common and are not mutually exclusive (Chory *et al.* 2000). Deploying large-scale data-gathering strategies for determining gene function in addition to smaller-scale hypothesis-driven research strategies is usually seen as a retrograde step by plant scientists, as the juggernaut of systematic screens appears to lack the finesse and focus derived from using accumulated knowledge to generate insight into biological processes. However, the fact that fewer than 10% of *Arabidopsis* genes have any experimentally defined functions (The *Arabidopsis* Genome Initiative 2000) demonstrates the immediate need for larger-scale generic approaches to ascribe rapidly and efficiently some degree of function to genes. Another benefit of systematic screens is the development of common descriptions of phenotypes, experimental conditions and stages of plant growth (Boyes *et al.* 2001). These provide a basis for comparison between phenotypes, microarray experiments, etc., necessary for data integration, described below.

One of the most important foundations for further advances in understanding plant gene function is the generation of lines with disruptions in all *Arabidopsis* genes. Due to the random nature of insertional mutagenesis, over 200 000 lines need to be made to achieve a more than 99% chance of finding an insertion in a 2 kb gene (Meissner *et al.* 1999). T-DNA and transposon insertional mutagenesis are both well-established methods and have been scaled up to produce several hundred thousand insertion lines. Table 1 shows some of the activities in this field. The distribution of reliable information and germ-plasm associated with gene insertions is a major challenge emerging from this work. The sequences of T-DNA and transposon insertion sites provide unequivocal evidence of disrupted genes, are searchable, and can be associated with genome sequence databases and with seed collections in public databases. T-DNA and transposon insertion sites are starting to be sequenced systematically (table 1) from TAIL and IPCR templates and deposited in public

databases. These activities will greatly benefit the rapid exploitation of the *Arabidopsis* genome sequence.

Currently, the most widespread gene tagging strategy uses insertion of elements such as transposons and T-DNA that lead to loss of gene function. Modified elements that alter gene expression upon insertion in or near a gene can provide important additional clues to gene function not apparent from insertions that lead to loss of gene function. Gene and enhancer traps (Springer 2000) reveal the tissue-specific expression patterns of disrupted genes or genes adjacent to an insertion, and can reveal disruptions in the heterozygous state. The thousands of gene trap lines now available (table 1) can be used to define systematically the spatial and temporal patterns of many *Arabidopsis* genes.

The integration of information about the cellular specificity of gene expression with data derived from microarray experiments will provide a key platform for interpreting gene function. For example, coexpression of unknown genes with genes of known function implies a direct or indirect association of gene function 'guilt by association' (Brown *et al.* 2000). Clustering and other learning approaches have been used to associate coregulated genes according to different conditions (e.g. circadian regulation) and treatments (pathogen responsive). The promoter sequences of coregulated genes were compared and conserved promoter elements were defined and functionally characterized. Pattern recognition approaches such as MEME (see <http://meme.sdsc.edu/meme/website/>) and ALIGNACE (Tavazoie *et al.* 1999) are powerful tools for defining promoter elements, but it is likely that *Arabidopsis*-specific solutions need to be developed and assessed to take account of plant-specific promoter architecture and known DNA-binding motifs. The increasing use of whole-genome gene expression methods will provide the background information needed to develop and refine models of gene regulatory circuitry, described below.

Mis-expression of genes, for example by insertion of a promoter-carrying T-DNA or transposon, is a powerful tool in the search for gene function. Phenotypes associated with loss-of-function of members of gene families that may have closely related, overlapping and 'redundant' functions may be very subtle, and mis-expression of a gene may provide clues relevant to the function of the native gene.

The subcellular location of gene products provides important clues to their function and the spatial organization of, and interactions between, cellular components are of central importance in cell biology. Antibodies coupled with fluorescent dyes are widely used to identify subcellular distribution of polysaccharide and protein epitopes, but these methods are relatively slow. The advent of reporter gene fusions and transgenic approaches provides the means for systematic large-scale testing of the subcellular distribution of proteins. In particular the GFP reporter protein has the potential to retain enzymatic activity as a fusion protein and can be sensitively detected in living cells at subcellular resolution using confocal microscopic imaging. The GUS reporter gene generally retains enzymatic activity as a fusion protein and can be sensitively detected. However, it cannot be localized in living cells, nor can it reveal subcellular locations apart from the nucleus, chloroplast and mitochondria. The GUS protein

Table 1. Gene disruption resources in *Arabidopsis*.

laboratory	insertion element	number	sequenced insertions
Salk Institute	T-DNA	70 000	12 000
Torrey Mesa Research Institute	T-DNA	100 000	100 000
Sainsbury Laboratory	DSpm	48 000	2000
Cold Spring Harbor Laboratory	Ds gene trap	30 000	10 000
Wisconsin	T-DNA	100 000	?
Genomanalysis im Biologische System Pflanze, Köln	T-DNA	70 000	5000

is enzymatically inactive after signal-peptide-mediated processing through the endomembrane system. Activity is recovered by tunicamycin treatment to inhibit protein glycosylation, which presumably inhibits GUS activity. Thus GUS protein fusions have significant potential, through gene trap transposons for example, for the systematic definition of the subcellular location of proteins in the endomembrane system as well as the nucleus, chloroplast and mitochondrion. GFP protein fusions, generated *de novo* or through gene disruption, can be used to reveal more precise localizations, for example sub-nuclear and other fine-grained structures in living cells. These surveys can be conducted on a genomic scale, using gene trap transposon lines for example, and will provide another critical dimension to our understanding of gene function.

One of the most interesting and distinguishing features of plants is their complex photoautotrophic metabolism, which elaborates a vast range of chemicals from simple components. Recently, major progress has been made in applying separation and detection technologies such as liquid- and high-pressure-liquid chromatography and mass spectrometric analysis to the analysis of plant metabolites (Roessner *et al.* 2001). The great sensitivity and speed of these analyses is suitable for high-throughput implementation, for example in forward screens to detect changes in metabolite composition. Nevertheless, the limitations of biochemical approaches still obtain: how can a change in metabolism be attributed directly to a change in or loss of protein function in the presence of other metabolic steps that may mask the proximal change? Biochemical genomics provides a more direct definition of the role of a gene in a biochemical pathway. In yeast, proteins were made as an array of glutathione synthase–open reading frame fusions and pooled (Martzen *et al.* 1999). Genes encoding particular biochemical activities were identified by assaying pools for a particular activity. The substrates used can be biochemical intermediates, targets for phosphorylation or other modification, or ligand binding. High-density arrays of proteins on a chip suggest the potential for miniaturizing and speeding up this process.

5. MAKING IT RUN

This review shows that diverse types of data, from gene expression to protein localization to metabolite analysis, can be obtained systematically in *Arabidopsis*. All of this information is relevant to gaining a complete understanding of the structure and function of cells at the molecular level, and it will accelerate progress in understanding the functions of individual genes. But this strategy will not

allow us to generate a comprehensive description of cellular processes at the molecular level. To achieve this objective we need first to capture the vast and diverse sets of data generated by systematic analyses, then to integrate this into databases, and then to use it to describe all known aspects of gene function in cellular and whole-organism processes. The final challenge is to represent this knowledge in an accessible and meaningful way. The solutions to these problems require novel bioinformatics solutions and a hitherto unachieved level of international coordination among plant scientists.

A relatively straightforward example of data integration illustrates the type of work remaining to be implemented on a genome-wide scale in *Arabidopsis*. It concerns establishing and testing models of gene regulatory networks. Three datasets—microarray data, reporter gene data showing the spatial patterns of gene expression and the sequences of promoters—need to be integrated. Promoter sequences associated with particular patterns of expression are candidates for regulatory elements, although their identification remains uncertain. Promoter sequences recognized by classes of transcription factors can also be used to establish regulons, although distinguishing among the members of large transcription factor families in *Arabidopsis* is a major hurdle (The *Arabidopsis* Genome Initiative 2000). Impressive progress has been made in defining circadian clock- and salicylic acid-modulated regulons using this strategy (Harmer *et al.* 2000). Modifying transcription factor function and examining the consequences for the spatial and temporal patterns of gene expression can establish, test and refine these gene regulatory networks. Machine learning computer algorithms can improve their performance using their previous experience and are thought to be especially promising for analysing high-dimensional data such as microarrays and images of gene expression, and reducing these data to fewer dimensions (for example by clustering) for visualization and further analysis (Mjolsness & DeCoste 2001). Hypothesis testing and refinement and statistical inferences are also mathematically supported by these methods.

Unprecedented opportunities have been presented to plant scientists, and in particular cell biologists, by completion of the *Arabidopsis* genome sequence and continued refinement of gene predictions. Many analytical tools and high-throughput systems will be generating a vast amount of diverse data that will need to be integrated and displayed to obtain detailed holistic views of plant cells and plant development. The challenge to scientists is to adopt to new ways of conducting biological investigation, using all the tools at hand.

REFERENCES

- Adams, M. D. (and 203 others) 2000 The genome sequence of *Drosophila melanogaster*. *Science* **282**, 2185–2195.
- Boyes, D. C., Zayed, A. M., Ascenzi, R., McCaskill, A. J., Hoffman, N. E., Davis, K. R. & Gorlach, J. 2001 Growth stage-based phenotypic analysis of *Arabidopsis*: a model for high throughput functional genomics in plants. *Plant Cell* **13**, 1499–1510.
- Brown, M. P. S., Grundy, W. N., Lin, D., Christianini, N., Sugnet, C. W., Furey, T. S., Ares Jr, M. & Haussler, D. 2000 Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA* **97**, 262–267.
- Chory, J. (and 13 others) 2000 Functional genomics and the virtual plant. A blueprint for understanding how plants are built and how to improve them. *Plant Physiol.* **123**, 423–425.
- Delmer, D. 1999 Cellulose synthesis: exciting times for a difficult field of study. *A. Rev. Plant Physiol. Plant Mol. Biol.* **50**, 245–276.
- Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanowski, A., Zollner, A. & Mewes, H. W. 2001 Functional and structural genomics using PEDANT. *Bioinformatics* **17**, 44–57.
- Hake, S. 2001 Plant biology: mobile protein signals cell fate. *Nature* **413**, 261–264.
- Harmer, S., Hogenesch, J. B., Straume, M., Chang, H.-S., Han, B., Zhu, T., Wang, X., Kreps, J. & Kay, S. 2000 Orchestrating transcription of key pathways in *Arabidopsis* by the circadian clock. *Science* **290**, 2110–2113.
- Hartwell, L., Hopfield, J. J., Leibler, S. & Murray, A. W. 1999 From molecular to modular cell biology. *Nature* **402**, C47–C50.
- International Human Genome Sequencing Consortium 2001 Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
- Kempin, S., Liljegen, S. J., Block, L. M., Rounsley, S. D., Yanofsky, M. F. & Lam, E. 1997 Targeted disruption in *Arabidopsis*. *Nature* **389**, 802–803.
- Lukashin, A. V. & Borodovsky, M. 1998 GENE MARK.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26**, 1107–1115.
- Martzen, M. R., McCraith, S. M., Spinelli, S. L., Torres, F. M., Fields, S., Grayhack, E. J. & Phizicky, E. M. 1999 A biochemical genomics approach for identifying genes by the activity of their products. *Science* **286**, 1153–1155.
- Meissner, R. (and 24 others) 1999 Function search in a large transcription factor gene family in *Arabidopsis*: assessing the potential of reverse genetics to identify insertional mutations in R2R3 MYB genes. *Plant Cell* **11**, 1827–1840.
- Meyerowitz, E. M. 1999 Plants, animals and the logic of development. *Trends Genet.* **15**, M65–M68.
- Mjolsness, E. & DeCoste, D. 2001 Machine learning for science: state of the art and future prospects. *Science* **293**, 2051–2055.
- Pavy, N., Rombauts, S., Dehais, P., Mathe, C., Ramana, D. V. V., Leroy, P. & Rouze, P. 1999 Evaluation of gene prediction software using a genomic dataset: application to *Arabidopsis thaliana*. *Bioinformatics* **15**, 887–900.
- Riley, M. 1993 Functions of gene products in *E. coli*. *Microbiol. Rev.* **57**, 862–952.
- Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzer, L. & Fernie, A. R. 2001 Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* **13**, 11–29.
- Springer, P. 2000 Gene traps: tools for plant development and genomics. *Plant Cell* **12**, 1007–1020.
- Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. & Church, G. M. 1999 Systematic determination of genetic network architecture. *Nature Genet.* **22**, 281–285.
- The *Arabidopsis* Genome Initiative 2000 Analysis of the genome sequence of *Arabidopsis thaliana*. *Nature* **408**, 796–815.
- Venter, J. C. (and 273 others) 2001 The sequence of the human genome. *Science* **291**, 1304–1351.
- Wagner, T. A. & Kohorn, B. 2001 Wall associated kinases are expressed through development and are required for cell expansion. *Plant Cell* **13**, 303–318.

GLOSSARY

- BAC: bacterial artificial chromosome
 GFP: green fluorescent protein
 GO: Gene Ontology
 GUS: β -glucuronidase
 IPCR: inverse polymerase chain reaction
 SMART: simple modular architecture research tool
 SNP: single nucleotide polymorphism
 TAIL: thermal asymmetric interlaced polymerase chain reaction
 TM: transmembrane loop