Mid-course evaluation of the Arabidopsis 2010 program

Executive summary

In 2000 the Arabidopsis community proposed an ambitious program to determine the function of every gene by 2010. This became the basis for the NSF AT2010 program, which has funded 86 projects in the first five years. The North American Arabidopsis Steering Committee held a workshop in Arlington, VA on Aug. 25 and 26, 2005, to evaluate the progress made toward the specific goals of the program and to recommend directions for the next five years. Prior to the meeting, input was solicited from the community through a web-based survey to which more than 580 researchers responded. Additional information on the impact of funded projects was obtained that described the number of stocks deposited, data generated and publications resulting from AT2010 projects. The workshop participants' assessment was that most of the goals for the first five years have been met or surpassed. Of particular note were the genome-wide resources including knockout lines and full-length cDNAs, which have been of remarkable utility to a large number of researchers. It was the participants' view that certain approaches toward functional analysis have had more impact than others. In particular, those that pioneered new approaches to understanding biological processes using high throughput and/or computational approaches have served as paradigms for other research efforts. In fact, it is expected that *Arabidopsis* will be the model for resource and tool development and application for all plants. For the remaining five years of the program the workshop participants recommend emphasis on the following areas:

1. **Benchmarking gene function**
2. **Developing genome-wide tools and reagents for analyzing gene function and regulation**
3. **Improving genome annotation and tools for visualization, annotation and curation**
4. **Improving database integration and developing new modeling and computational tools**
5. **Exploring exemplary networks and systems**
6. **Analyzing non-protein coding genes**
7. **Leveraging natural variation to understand gene function in *Arabidopsis thaliana***
8. **Localizing gene products at the cellular and subcellular level**
9. **Facilitating metabolomics and ionomics**
10. **Engaging the broader community**
11. **Enhancing international collaboration**

The workshop also looked beyond 2010 to challenges that could form the basis for an Arabidopsis 2020 program. This ongoing and future research program should have critical impacts on many areas of basic science, agriculture, engineering and environmental improvement as well as on all aspects of plant biology.

Progress towards 2010 goals

The first five years of the Arabidopsis 2010 Project have been quite successful, achieving and in some cases exceeding many of its objectives. The 2010 Project has yielded valuable resources and tools for functional genomics. There have been some important advances in understanding gene function through the analysis of specific processes. An integral component of the 2010 Project has been the promotion of collaborations both national and international, and as such, the 2010 project has served as a cornerstone of the Multinational Coordinated *Arabidopsis* Functional Genomics Project. As a whole, the Arabidopsis community has achieved many of the benchmarks put forth by the Salk Institute 2010 Workshop held in January 2000 (http://arabidopsis.org/info/workshop2010.jsp).

In the first five years, critical advances have been enabled by resource development projects, comprising 21% of 2010 awards (See Figure 1 for distribution of awards). These projects have had a large impact in the field by vastly expanding the resources and tools that are available for ascertaining gene function, enabling investigators throughout the world to solve a wide range of biological problems (The Multinational Coordinated *Arabidopsis thaliana* Functional Genomics Project, Annual Report 2005). Indeed, a landmark achievement of the 2010 Project has been the generation of T-DNA/transposon collections of 26,000 sequence-indexed gene insertion mutants, all of which have been made publicly available as seed stocks. By a conservative estimate, these collections provide potential nulls (insertions within exons) for ~70% of the genes in Arabidopsis. This number substantially contributes to the total statistic of ~92% of Arabidopsis genes carrying a T-DNA or transposon insertion within their transcription unit, including the T-DNA knockouts provided by the Arabidopsis knockout facility. Currently, homozygous lines are being developed and verified for T-DNA insertion alleles of 25,000 genes. The TILLING lines provide yet another invaluable resource for reverse genetics. To date, approximately 5800 mutations (EMS-generated) have been reported through TILLING. The Arabidopsis TILLING project has served as a paradigm for similar projects in other organisms. Other vital resources are full-length cDNA and open-reading frame (ORF) clones, currently representing 70% of known genes. This collection, which is publicly available, greatly facilitates genome annotation as well as protein analyses. RNAi knock-down lines are currently being developed for 20,000 genes by AGRIKOLA, an EU funded project. The 2010 Project has also contributed to the characterization of a set of ecotypes, establishment of recombinant inbred (RI) lines and identification of a large collection of SNPs. There are several RI lines available, and many additional sets of RI lines are being made. All of these resources are having a tremendous impact on the rate at which functional genomics is proceeding.

Approximately, nine percent of 2010 projects have been used to develop core resources, including bioinformatics/database resources such as TAIR, which

support many community-wide advances. Genome-wide sets of gene-specific probes for expression analysis are available, and numerous expression profiling datasets have been deposited in public databases. A major milestone is the availability of the AtGen Express reference transcriptome, catalogs of small RNAs and tiling array transcription maps. Although, not all directly funded by the 2010 Project, it has contributed synergistically to all of these developments, including improvements in whole genome annotations.

The remaining 70% of the 2010 projects have focused on achieving a greater understanding of gene function and processes. These projects have resulted in notable advances in certain areas, including pathways and networks, biotic interactions, small molecules (ionome), clone-based proteomics, gene regulation, and growth and development, as evidenced by a large number of publications in high impact journals and database contributions. Certain approaches toward functional analysis have had more impact than others. In particular, those that pioneered new approaches to understanding biological processes using high throughput and/or computational approaches have served as paradigms for other research efforts. Efforts that focused on elucidating the function of specific gene families have generally had lower impact (although there are some notable exceptions).

It is clear that improved technologies will continue to accelerate our understanding of gene function in Arabidopsis. This can be best served by the 2010 Project through emphasis on the further development of resources and new technologies, effectively empowering the entire Arabidopsis community to advance the goal of understanding gene function through genome-wide studies. At the same time, there is unquestionable value in continuing to utilize existing technology at production scale to provide a base level description of every gene.

In addition to the scientific progress described above, the first five years of the 2010 project have been marked by increasing international cooperation and exchange among Arabidopsis researchers. Highlights have been the International Congresses on Arabidopsis Research, held twice in Europe and three times in the US in the past five years, joint grant panels between German and US funding agencies and a graduate student exchange program between German and US labs. The Multinational Arabidopsis Steering Committee, with its full-time coordinator has played an important role in implementing and documenting international collaborative efforts among Arabidopsis researchers.

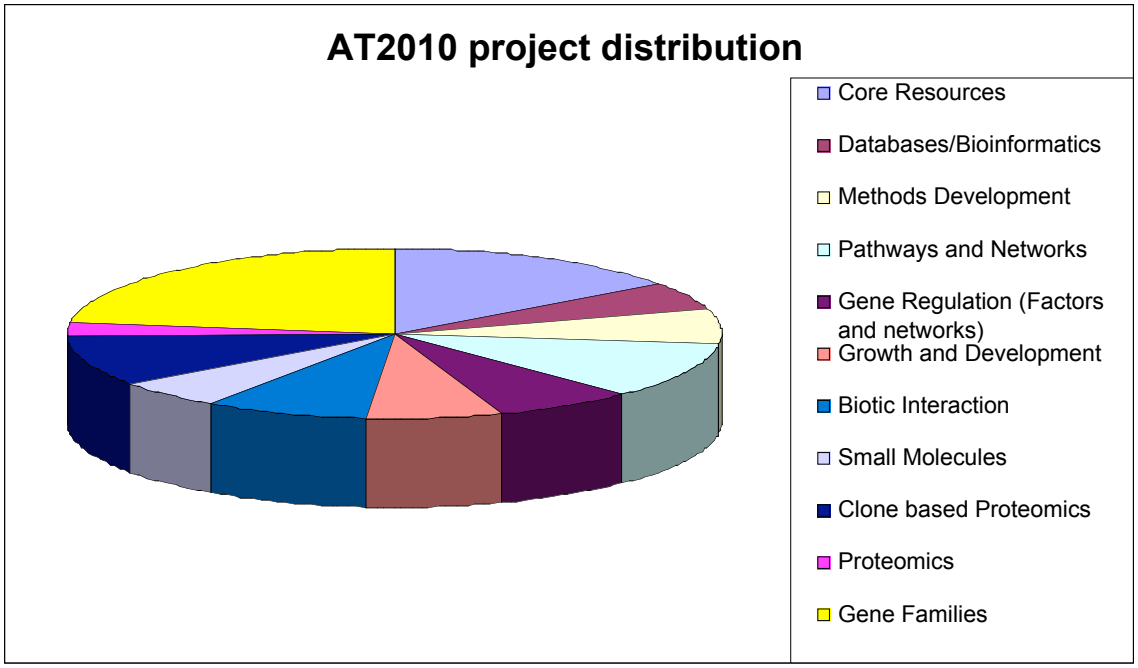

Figure 1.: AT2010 award distribution

Recommendations for the next five years.

1) Benchmarking gene function. The stated goal of the Arabidopsis 2010 Project is to determine the function of every gene. This requires that certain benchmarks for assigning gene function be defined and applied genome wide. The underlying rationale is that significant information about the functions of both characterized and uncharacterized genes can be revealed from the following data types: expression patterns, interacting partners and cis-elements. Although detailed information about the functions of many genes will continue to be generated over the next 5 years, it should be possible by the completion of this project to define each gene by the following criteria:

a. Gene expression data that includes information on temporal and spatial patterns of transcription under different environmental conditions and in specified genotypic backgrounds. Development and adherence to community-wide standards for collection and presentation of expression data are highly encouraged. We recommend that large-scale datasets be made available in a format that enables rapid comparison and effective utilization of reproducible information.

b. Identity of interacting protein partners. In a number of model organisms and humans, high-throughout studies of protein-protein interactions have contributed valuable information on gene function. Two hybrid studies, affinity purification/mass spectrometry, and proteome chip experiments have generated vast amounts of interaction data. To facilitate

completion of the Arabidopsis 2010 Project, we encourage continued development and application of high throughput protein-protein interaction studies that are both comprehensive and cost effective. Multiple approaches that include appropriate blending of in vivo and in vitro interaction studies should be supported. It is important that measures of specificity and sensitivity be reported and that quantitative information about signal strength and confidence be included.

c. Identity of cis-regulatory elements revealed in part through comparative phylogenetic approaches and the global mapping of regulatory elements and factors.

2) Developing genome-wide tools and reagents for analyzing gene function and regulation. An important goal of the 2010 project is to maximize the impact on both plant research and our understanding of plant biology by leveraging genomic information with community-wide reagents and tools. Many different types of resources are expected to be widely utilized by the plant community and enhance research in both *Arabidopsis* and other plants. In fact, it is expected that *Arabidopsis* will be the model for resource and tool development and application for all plants. Development of tools and approaches that provide quantitative readouts, are cost effective and comprehensive and can be readily adopted by the scientific community should be emphasized. We strongly recommend development of the following resources and tools:

a) *Mutant collections for phenotyping.* Phenotypic information is expected to have enormous impact on our understanding of gene function and biological processes. The large insertion collections that have been generated are already producing useful data. We recommend support for efficient and inexpensive distribution of comprehensive sets of these mutant collections. We also support projects that would efficiently and inexpensively complete the insertion collection and/or develop methods for high throughput phenotyping. In addition, the development of a standardized ontology for quantifying and comparing phenotypes is essential for this project. It is expected that researchers will carry out genome-wide screens with these collections, and we recommend a centralized database for users to deposit their results.

b) *Other reagents for phenotyping*. In addition to the insertion collection we recommend the production of other collections that support acquisition of mutant phenotypes. These include completion of an ongoing RNAi collection against individual genes and gene families, construction of a conditionally expressed RNAi collection, and collections of other lines in which genes are conditionally active.

c) *Genome-wide reagents*. Completion of the full-length cDNA and ORF collection and production of other collections that would enhance *Arabidopsis* research such as collections that overexpress each *Arabidopsis* protein.

d) *Pilot studies*. We recommend pilot studies for production of affinity reagents (e.g. antibodies, aptamers) to *Arabidopsis* proteins (for immunoprecipitation, localization, immunoblots etc) and efficient epitope tagging of proteins expressed at endogenous levels (see below Section 8).
e) *Tools.* We strongly recommend the development of genome-wide tools that would facilitate *Arabidopsis* research. Ideally these tools should be developed and rapidly assimilated into the plant community (not just *Arabidopsis*). In many cases, pilot studies may be important to test feasibility for genome-wide applicability. Examples of tools that will be valuable to the community are methods for inactivating redundant genes to reveal phenotypes, homologous recombination, global methods for analyzing posttranslational modifications, protein profiling, analyzing protein-small molecule interactions, subcellular localization, transcription factor binding and epigenetic phenomena (see below Section 8).

3) Improving genome annotation and tools for visualization, annotation and curation. Arabidopsis is the leading reference plant, therefore optimal functional annotation is required for the plant sciences community. Annotation is a dynamic process, capturing the advances of experimental work. As such, the workshop participants concluded there was a need for continued improvement of genome annotation, and development and implementation of annotation and visualization tools. In addition, it is recommended that the frequency of whole genome annotation updates should increase. Updates should consist of both semi-automated and manual curation. The annotation should consist of both sequence-based annotations and non-sequenced-based annotations using controlled syntax consistent with efforts of other model organisms. Additional sequence development and experimental approaches should complement existing resources. Examples of sequence-based resources could include more full-length cDNA development, surveying genomic sequence of phylogenetic node species, and surveying sequences of *A. thaliana* ecotypes (see also section on natural variation). As additional advances are made in computational and experimental approaches, these should be applied to the sequence. This would include, but not be limited to, identification of SNPs, 5' and 3' ends, regulatory elements and protein motifs. Non-sequence based annotation should include the integration of experimental evidence of function available from both literature and submission by the community. Examples of non-sequence based evidence of function would include, but not be limited to, results from expression, ChIP/chip, genome tiling arrays, localization studies, phenotypic analysis, and interaction data. Priorities include development of standardized protocols and tools to allow for submission of common data types for genome annotation and development or adoption of visualization tools for viewing and linking the sequence annotations, in addition to new means of data integration. For example, it would be desirable to have a browser that would allow users to view genomic annotations in the context of

multiple annotation types and would act as a portal to the primary data providers.

4) Improving database integration and developing new modeling and computational tools

Database and data integration

The vast amount of genomic data generated during the first 5 years of the 2010 Project and the anticipated community resources forthcoming during the second half of 2010 necessitate substantial development of new and improved computational tools for data storage, visualization and integration. TAIR is a model of integration of information for streamlined access, but it is one of many nodes in the 2010 cyber network. A seamless cyber infrastructure providing a portal to a sustainable distributed network of resources is essential for the *Arabidopsis* community and the broader scientific community.

Data integration of phenotype and ontology resources is a high priority for the second phase of 2010 that will require the development and community-wide implementation of more sophisticated and standardized methods of phenotype quantification, visualization and databasing of information.

Modeling and computational tools that are tightly integrated with experimental approaches

A range of modeling and computational approaches are emerging in concert with the experimental work. A range of quantitative modeling approaches and simulations should be used to analyze and integrate information from microarray analyses, global protein profiling, gene networks, cell-cell communication networks, metabolic regulatory networks, and multi-level analyses. The most effective approaches to these modeling efforts will be embedded in experimental work and validated.

5) Exploring exemplary networks and systems. The original 2010 project aimed to determine the function of every gene. During the first phase of the program it has become increasingly clear that individual genes may serve multiple functions, and that multiple genes work together to regulate a variety of biological processes. In the second half of the 2010 program, projects that pioneer new approaches to the integration of omic information into networks will provide a bridge between genomics and systems biology. Using subsets of information about nodes and edges that define genome-wide interactions, a variety of useful biological networks can be constructed. These include but are not limited to: developmental, biotic, physiological and metabolic processes – as well as meta-networks connecting these processes. These studies should enable the discovery of functional information associated with networks and different levels of regulatory information leading to the description of detailed regulatory circuits controlling plant growth and development. Such studies should lead to the identification of regulatory hubs, which ultimately may be altered to modulate plant growth, yield and productivity.

6) Analyzing non-protein coding genes. The last five years have witnessed a remarkable expansion in our understanding of how non-protein coding RNAs play a critical regulatory role in eukaryotic cells and more specifically, in plants. These molecules include but are not limited to: miRNAs, siRNAs, stRNAs, snoRNAs. This nascent understanding of the biology of non-protein coding RNAs has enabled significant technological advances such as routine use of RNAi constructs to generate functional gene knock-outs. Nonetheless, the genes encoding non-protein coding RNAs have been significantly under-represented in genome-wide expression analysis, regulatory network analysis and annotation. Functional analysis, computational modeling and community-wide tool development based on these classes of genes should be a focus in the second half of the 2010 project.

7) Leveraging natural variation to understand gene function in *Arabidopsis thaliana.* Our understanding of gene function in *Arabidopsis thaliana* will be greatly enhanced by comparative genomic information from related species, as well as population genomic studies of natural variation within and between informative ecotypes. It will be essential to identify and sequence species at informative phylogenetic nodes***.*** Comparative genomic analysis of these species can identify both conserved and rapidly evolving regions of the *A. thaliana* genome. Such information will be vital for whole-genome functional analysis. For example, comparative genomic data can be used to identify conserved cis-regulatory elements as well as coding regions under strong functional constraints. Comparative analysis of sequence evolution and gene loss following duplication events will shed light on whole-genome patterns of gene redundancy, subfunctionalization, and neofunctionalization.

It will be equally essential to develop tools for whole-genome population biology. Analysis of natural variation within and between informative *A. thaliana* ecotypes will be extremely valuable for understanding gene regulation and function from the level of gene sequence to whole plant performance. To this end, it will be important to complete the ongoing development and wide distribution of informative sets of RILs and ecotypes to the community for functional studies. Genome sequencing within and between appropriately chosen ecotypes or natural populations will also be valuable. This will allow whole genome molecular population genetic studies to identify unusually conserved or rapidly evolving genes for functional studies, as well as SNP development for analysis of natural variation and population structure. Such studies will permit the linking of allelic variation to function through linkage disequilibrium mapping. In addition, studies of natural variation in whole genome expression will help to elucidate patterns of *cis* and *trans* acting regulation.

8) Localizing gene products at the cellular and subcellular level. Wiring diagrams obtained by mapping the interactome network and

characterizing the function of its components need to be complemented by genome-wide information on dynamic aspects of expression. Information will be needed about gene and protein expression patterns at the organ, tissue, cellular and/or subcellular levels in the living plant. Localizome mapping consists of systematically obtaining information on where (in what cells or tissues) and when (at what stage(s) of development and/or under what conditions) large numbers of genes are expressed, and where and when their products are localized at the subcellular level.

Genome-wide localization of expression projects include, but are not limited to, the generation of transgenic plants carrying appropriate recombinant gene constructs such as promoters fused to sequences encoding fluorescent marker proteins (e.g. GFP), promoters fused to protein-encoding open reading frames (ORFs), themselves fused to sequences encoding fluorescent marker proteins, ORF sequences fused to 5' and/or 3' untranslated regions (UTR), etc. Such approaches will require the production of genome-wide flexible cloned promoter, ORF, 5' and 3' UTR resources. Such efforts should ideally be compatible with other approaches described in this Program, such as the mapping of interactome, transcriptional and other networks and the development of increasingly sophisticated genome-wide RNAi methodologies. Importantly, methods for automated detection and visualization of expression patterns should be developed, as well as new modeling strategies that integrate the localizome data with interactome network models.

Although amenable to automation and therefore usually preferable for genome-wide determination of localization of expression patterns, transgenic methodologies can be complicated by sensitivity, specificity, and dynamic range issues. Particular attention should be devoted in any localizome project to determine the overall data quality and the extent of false positive and false negative problems.

9) Facilitating metabolomics and ionomics. The production, regulation and function(s) of small molecules such as metabolites as well as the transport and activities of ions have important roles in biology. Knowledge of small molecule biology in *Arabidopsis* is not only fundamental for a holistic understanding of plant biology, but also in the context of plants as a source of food for both humans and animals as well as sources of pharmaceuticals.

Important advances have been made in the context of the Arabidopsis 2010 program in the development of high-throughput quantitative analyses of ions. In contrast, robust methods for rapidly detecting, quantifying and determining the activities of large numbers of metabolites are limited. We support development of tools and pilot studies that globally monitor small molecule profiles, transport, activities and regulation.

10) Engaging the broader community. Empowering the broader community by providing genomic resources continues to be a major focus of the 2010 effort. Researchers at universities, undergraduate institutions, and

community colleges have and will continue to benefit from the information, tools, and other resources generated through 2010. To maximize the benefit, developing the cyber infrastructure for the dispersed network of 2010 resources that continues to grow beyond the scope of any single database is imperative. Timely data and tool sharing will increase the value of 2010 funded work to the research community.

As genomics finds its way into the pre-college curriculum and the social and policy decision-making arenas, 2010 investigators are encouraged to explore ways to reach these audiences. Access to information about 2010 research is one gateway for this larger audience and should be considered in the further development of databases, data integration, and cyber infrastructure. Actively engaging undergraduates, high school teachers, and high school students in appropriate aspects of the research is another avenue that has been successful in the first five years of 2010. A specific emphasis on reaching under-represented groups is important in increasing their participation in genomics research and building scientific workforce capacity.

Graduate student and postdoctoral training in plant genomics will continue to be enhanced by 2010 projects. Trainees will have opportunities to participate in cutting edge genomics research, including informatics, and will benefit from the resources generated when they move on to more independent positions. The next five years offer a critical window to build researcher capacity maximizing the benefits of the 2010 research.

While the focus of 2010 is on the *Arabidopsis* genome, this reference plant will leverage work in other plant genomics communities as well as in communities working on other organisms. The 2010 program offers an opportunity for the *Arabidopsis* genomics community to interface with other plant genomics communities, and with communities working on more diverse organisms.

11) Enhancing International collaboration. A major reason for success in the establishment of Arabidopsis as a reference plant has been the strong international collaboration that has developed. International collaboration is particularly important in genomics research, as this scientific revolution requires costly efforts to generate resources and tools for genome-wide analyses. Mechanisms should be sought to avoid unnecessary duplication of work. A paradigmatic example of success associated with international collaboration has been the completion of the sequence of the Arabidopsis genome that has been publicly available for the world community of scientists since 2000.

International collaboration should be further stimulated so that researchers share all seed and genomic resources. These should be maintained in public stock centers making Arabidopsis research readily accessible to the worldwide community. Equally important is the need for international

collaboration on the development of a common database or a confederated database system, which stores and makes accessible the vast array of genomic and phenotypic data on Arabidopsis.

Continued scientific exchange at the level of workshops, international conferences and extended visits by senior researchers, post-docs and graduate students is essential to maintain the high level of cooperation and knowledge dissemination among Arabidopsis researchers. An excellent example is the recently instituted graduate exchange program sponsored by NSF between German and US labs. Increasing the frequency and depth of exchanges with other disciplines, particularly those with a strong quantitative basis, will also be critical for attaining the goals of the 2010 program.

General issues

Measurements and Quality Control

Genome-wide projects generate large amounts of data. However the data are most valuable when quantitative measurements are made and the quality of the measurements is ascertained. We recommend that cost effective technical and biological replicates be performed. Biological replicates in particular ensure that the data collected are reproducible. Furthermore, it is essential that signal strength and confidence values be associated with each data point. In addition, studies and methods should also report measurements of sensitivity (false negative) and specificity (accuracy) ideally as determined using alternative methods to validate results. (Note: it is important to ascertain quality measurements at three stages: tool development, pilot studies and during large scale projects. Validation should also be performed by each group whenever multiple groups are involved).

Deliverables

In the first 5 years of its existence, the Arabidopsis 2010 project has generated broad knowledge about plant function, accompanied by the development of biological reagents, software and experimental tools. Existing policy clearly dictates that biological and sequence-based reagents should be deposited in defined common repositories, such as the ABRC and GenBank, respectively. Unlike these examples, it is less clear where other data types should be made available to the public because this was not well defined by the 2010 project. As a result, much information that could have added value for annotation of the genome has not been integrated into existing resources. This situation should be addressed for the second part of the 2010 project. Publication of findings is no longer considered sufficient to build the desired comprehensive, integrated resource for the community. Therefore it is strongly recommended by the workshop participants that policies should be developed to guide deposition of all project deliverables in public repositories. For each one of the deliverables on a project, where possible, international repositories should be defined and a time-line for release to the community should be stipulated. One such example of a policy is the Bermuda Standard, which covers sequence information. For example,

SNPs could be sent to dbSNP and array results could go to GEO. This would allow individuals and community-based resources to retrieve data in a standardized format. Where no international repositories exist, a community-based resource should be identified, where available. One example of this would be the functional annotation of proteins, which could be communicated to TAIR to improve the genome annotation. Though project web sites should not be a substitute for deposition in a long-term repository, they are still desirable. However, for impact and consistency, consideration should be given as to what should be the minimum content for these sites, possibly including research goals and progress, participants and publications. The responsibility for capturing and integration of data sets should be shared between the data provider and the community resource. As new 2010 projects are established sufficient funds should be budgeted to meet this goal.

Towards an Arabidopsis 2020 Project

We envision an Arabidopsis 2020 Project that enables the international community of plant biologists to analyze, understand, and manipulate the full spectrum of biological processes required to make a plant that functions effectively and predictably under both standardized laboratory conditions and complex natural environments. It is likely that by 2010 our understanding of gene function and networks will be primarily in their static state. For 2020, a major objective will be to understand the dynamic properties of the genes and networks that control plant functions. Another aim will be to understand how plants differ in their use of these genes and networks within populations and between species. The analytical methods and computational tools required to complete this project will engage scientists with diverse interests ranging from engineering and computer science to systems biology and ecological genomics. The benefit to society will be measured by our enhanced ability to define, explain and modulate a wide range of fundamental processes unique to plants and ultimately linked to continued advances in agriculture, human health, energy utilization, and environmental preservation.

Participants:

Philip Benfey (Duke University)
Caren Chang (University of Maryland)
Gregory Copenhaver (University of North Carolina at Chapel Hill)
Gloria Coruzzi (New York University)
Liz Dennis (CSIRO Plant Industry Australia)
Joe Ecker (Salk Institute)
David Meinke (Oklahoma State University)
Elliot Meyerowitz (California Institute of Technology)
Javier Paz-Ares (Centro Nacional de Biotecnologia-CSIC, Madrid)
Annie Schmitt (Brown University)
Susan Singer (Carlton College)
Mike Snyder (Yale University)

Kate VandenBosch (University of Minnesota)
Marc Vidal (Harvard University)
Doreen Ware (Cold Spring Harbor Laboratory)