THE MULTINATIONAL COORDINATED *ARABIDOPSIS THALIANA* FUNCTIONAL GENOMICS PROJECT

Project Goals

In 1990, a set of research priorities and goals were put forth which would lead to the knowledge of a complete plant genome sequence. With that goal successfully met, the time has come to put forth a new set of goals.

In this respect, probably the most important information provided through the analysis of the *Arabidopsis* genome sequence was the discovery of the limits of our current understanding of plant gene function and of the roles that the genes play in the multiplicity of processes involved in plant metabolism, development and interaction with the environment. The several decades of pre-genome research in *Arabidopsis* has yielded experimental data on less than 10% of *Arabidopsis* genes. It is the task of the *Arabidopsis* research community to ensure that in the same timely and cooperative manner as we approached the challenge of sequencing the genome, we use the genome information to understand the function of all *Arabidopsis* genes and in this way to achieve comprehensive knowledge of plant biology.

In the genome sequencing era, the many groups making up the *Arabidopsis* genomics community were working toward a single goal. Technologies and data converged on the single endpoint of a sequenced genome. The destination of a sequenced genome has now become a launchpad; from this launchpad will spring many technologies and types of data with which we will move forward to the new, multi-pronged goal of complete functional knowledge of an *Arabidopsis* plant.

The Multinational Coordinated *Arabidopsis* Functional Genomics Project is an idea that developed from a workshop that was held in early 2000 entitled "Functional Genomics and the Virtual Plant: A blueprint for understanding how plants are built and how to improve them" (text available at http://www.arabidopsis.org/info/workshop2010.html). Resulting from the workshop were new objectives for the *Arabidopsis* community, "to exploit the revolution in plant genomics by understanding the function of all genes of a reference species within their cellular, organismal and evolutionary context by the year 2010." The details of this project can be viewed as the second phase of the far-reaching vision described by the scientists who launched the Multinational Coordinated *Arabidopsis thaliana* Genome Research Project in 1990.

To achieve a complete understanding of the biology of a plant, we must in essence create a wiring diagram of a plant throughout its entire life cycle: from germinating seed to production of the next generation of seeds in mature flowers. These processes are controlled by genes and the proteins they encode. They are directed by both intrinsic developmental cues and environmental signals. The long-term goal for plant biology following complete sequencing of the *Arabidopsis* genome is to understand every molecular interaction in every cell throughout a plant lifecycle.

The ultimate expression of our goal is nothing short of a virtual plant which one could observe growing on a computer screen, being able to impose environmental changes and to stop this process at any point in that development, and with the click of a computer mouse, accessing all the genetic information expressed in any organ or cell and the molecular processes mediated by these factors.

Complete knowledge of the workings of a plant – a "virtual plant" – will allow a profound understanding of the biochemical processes and physiological responses of a plant. This knowledge will allow hypothesis testing and

experimentation leading to the modification and improvement of crops. It will result in a future in which we can limit our dependence on chemical pesticides and fertilizers, lessen our negative impact on the earth, and maximize crop yields to feed a growing world.

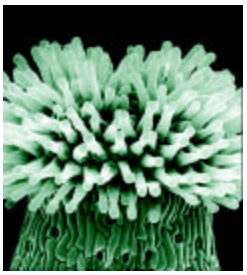
Scientific Objectives

The objectives now being put forth for the world-wide *Arabidopsis* research community include the development of expanded genetic toolkits as a service to the research community, implementation of a whole-systems approach to the identification of gene function from the molecular to evolutionary levels, expansion of the role for bioinformatics, development of human resources, and international collaboration.

1) An Expanded Genetic Toolkit

A key strength of *Arabidopsis* as a model is its facile forward genetics, largely due to its relatively small size and short life cycle. One can isolate mutants disrupted in many processes and study the effects of each mutation. Despite this, roughly 40% of the genes found in the genomic sequence do not encode a protein of predictable function.

Thus, out of the 25,500 predicted genes in the *Arabidopsis* genome, ca. 10,000 have a sequence that tells us nothing about what they do. For instance, sequence reveals that there are about 1,200 protein kinases in the *Arabidopsis* genome; however, to date, the *in situ* functions of only about a dozen of these have been found by forward genetics. Similarly, we know the identity of the ligand that binds the putative receptor site in only a handful of the approximately 600 members of the receptor kinase subfamily.



Since forward genetics relies on a phenotype arising from a single gene mutation, it is likely that a large number of genes may not be easily characterized using this approach. In order to identify functions for these genes, we need to develop a more sophisticated genetic toolkit for both forward and reverse genetic screens.

Overexpression of natural or altered proteins can provide insights into families of genes that are collectively essential. A straightforward, albeit laborious, approach that resembles strategic breeding weds reverse genetics and forward genetics. In this approach, the genome sequence is used to locate protein family members. Knockout mutations are located for all the genes in a family and the lines are crossed so that one plant contains knockout alleles of all the members of the gene family, creating a more robust phenotype. A process like this one can lead to increased understanding of the functions of genes that exist in gene families, and is not possible without the entire sequence in hand.

Short-term goals:

- Comprehensive sets of sequence indexed mutants, accessible via database search
- Whole genome mapping procedures
- Facile conditional expression systems for sensitized and saturation screens for rare alleles

Mid-term goals:

- Development of many combinations of large families of Recombinant Inbred Lines and Genetic Substitution Lines to allow facile analysis of natural variation
- Construction of comprehensive sets of defined deletions of linked, duplicated genes
- Development of methods for directed mutations and site specific recombination
- Establishment of libraries (complete collections) of transgenic lines for short term overexpression or repression of gene function

Goals for 2010:

- Plant artificial chromosomes
- Approaches to allow directed combination of multiple genetic modifications (such as strategic breeding)

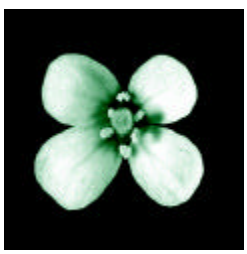
2) Whole-Systems Identification of Gene Function

The post-genome-sequence era allows a shift from single-gene or single-process research to whole-systems approaches to understanding plant biology. Identified as tools for the goal of global understanding of the plant are global analysis of gene expression, global analysis of protein dynamics, metabolite dynamics, global catalogues of molecular interactions, and comparative genomics.

One of the most critical aspects of this project is the enabling technologies that must be developed to achieve the scientific objectives we are putting forth. The proposed research will lead to a new array of fully developed technologies for scientific investigation, particularly in the areas of proteomics and metabolomics. While these technologies will be developed for the purpose of *Arabidopsis* research, they will naturally spill over into the world of research on economically important plant species and indeed research on all complex systems.

Global analysis of gene expression

An understanding of gene function begins with knowledge of when and where each gene is expressed during the normal development of a plant. Taken together, this information will become a platform from which the concerted action of gene sets in the formation of tissues and organs can be elucidated. Further, examination of the changes in gene expression that occur with environmental changes will illustrate the dynamic nature of gene regulation in plants.



Short-term goals:

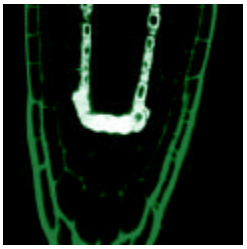
- Construct gene-specific DNA probes for expression analysis at high sensitivity and wide dynamic range
- Define full-length cDNAs to facilitate annotation of the genome and subsequent analyses of protein expression

Mid-term goal:

- Determine global mRNA expression profiles at the organ, cellular and subcellular levels under a wide variety of precisely defined (by quantitative measures) environmental conditions

Goals for 2010:

- Identify the cis-regulatory sequences of all genes
- Determine the regulatory circuits controlled by each transcription factor in the genome
- Uncover the *in planta* role of every gene through forward or reverse genetic approaches



Global analysis of the plant proteome

The sum of gene expression changes is translated through development into the proteins from which cellular machines are built. Understanding protein dynamics will enable prediction of what machines exist and how they work throughout a plant's life cycle. This aspect of *Arabidopsis* research has become especially important, because recent experimentation suggests that RNA changes alone are remarkably poor predictors for final changes in protein levels or enzymatic activity.

Short-term goals:

- Develop facile technology for heterologous expression of all proteins
- Produce antibodies against, or epitope tags on, all deducted proteins
- Catalogue protein profiles at organ, cellular and subcellular levels under a wide variety of environmental conditions

Mid-term goals:

- Achieve a global understanding of post-translational modification
- Define the subsets of genes encoding small peptides or small RNAs and their gene products, which will require the development of novel analytical tools for analysis and modulation of expression

Goals for 2010:

- Identify biochemical functions for every protein
- Determine three-dimensional structures of representative members of every plant-specific protein family

Global analysis of metabolite dynamics

Plant growth and development is dictated, to a large degree, by the uptake, trafficking, storage and use of low molecular weight metabolites. Plant cellular factories produce a bewildering array of secondary metabolites, upon which a large amount of drug and product discovery are based. Understanding metabolite dynamics will result in more efficient use of soil and water based nutrients and will allow rationally designed food and pharmaceutical production in plant factories.

New, sophisticated mass spectrometry techniques provide the *Arabidopsis* researcher a different type of microscope: one that allows us to monitor potentially thousands or tens of thousands of small molecules at a time. Using these "microscopes", the mass spectrometers will allow us to see changes at the molecular level that the eye cannot detect at the morphological level. This will be critical for connecting genetic changes with changes in the expression of enzymes and the metabolic pathways they comprise. Much like we collect RNA data using DNA chip technology, we need to expand our phenotypic screens to include newly developed

and expensive instruments that open our eyes to the fascinating and complex chemical world of plants.

Mid-term goal:

- Global metabolic profiling at organ, cellular, subcellular levels under a wide variety of precisely defined (by quantitative measures) environmental conditions and in many different genetic backgrounds

Goal for 2010:

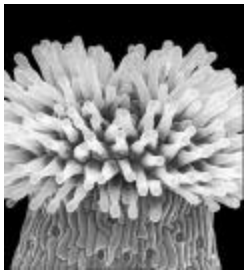
- Whole systems analysis of the uptake, transport and storage of ions and metabolites

Global catalogues of molecular interactions

The ultimate arbiters of cellular function are the complex protein machines encoded by the mRNA population in each cell at any time during development. Ultimate understanding of the cellular mechanics of a plant requires a catalogue of molecular interactions that occur in each cell of the organism throughout its lifecycle. This ambitious experimental layer incorporates an understanding of the gene expression, protein and metabolite dynamics of the plant.

Goal for 2010:

- Achieve a global description of protein-protein, protein-nucleic acid, protein-metal, and protein-small molecule interactions at organ, cellular, subcellular levels under a wide variety of precisely defined (by quantitative measures) environmental conditions.



Comparative genomics

Completion of the *Arabidopsis* genome sequence provides significant leverage for future plant genome projects. The reference genome is a platform from which useful comparisons are simplified. We will ultimately be able to predict the evolution of new gene function by comparative genomics. We can glimpse the power of comparative genomics as a tool to understand plant evolution and diversification through the recent strides made in the understanding of plant disease resistance gene structure.

Therefore, as a part of functional genomics work centered on *Arabidopsis*, work should be done on other species to enable the comparative genomics that will give us the power to apply the knowledge gained from this initiative to crop production.

Short- and Mid-term goal:

- Identify species for survey genomic sequencing based on an expanded definition of phylogenetic nodes

Goals for 2010:

- Survey genomic sequencing and deep EST sampling or sequencing of gene-rich regions from phylogenetic node species
- Define a predictive basis for conservation versus diversification of gene function
- Complete within-species genomic sequence comparisons
- Develop tools for whole genome population biology

3) An Expanding Role for Bioinformatics

Achieving the above goals will require significant investment in and development of bioinformatics tools and databases from which the information required to build the virtual plant will be stored and extracted. A significant effort in this area must be expended in close coordination with the biological aspects of the project.

Ultimately, the database that we envision will provide a common vocabulary, visualization tools, and information retrieval mechanisms that permit integration of all knowledge about an organism into a seamless whole that can be queried from any perspective. Of equal importance for plant biologists, an ideal database will permit scientists to use information about one organism to develop hypotheses about other, less well-studied organisms. Thus, our goal should be to develop facile tools that permit an individual working outside the model species to formulate a query based on the organism of interest, have that query directed to the relevant knowledge for the plant models, and present the information about the models in a way that can be understood by the plant biology community at large.

Database architecture allowing easy integration with other databases will be an essential component of this effort. Divergent types of data (e.g. expression array data and in situ hybridization, but also precise information about experimental setup and growth conditions defined by quantitative measures) will need to be integrated and archived. The ability to generate these datasets will easily outpace the ability to rationally maintain, manage, and extract utility from this data. Hence, there is a critical need to invest in novel data-mining approaches and to also bolster support for current databases.

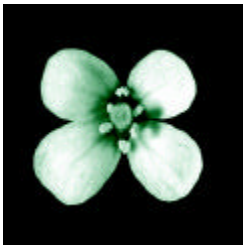
Ongoing goals:

- Develop new cellular and whole plant visualization tools
- Attract bioinformatics professionals to direct and aid in database creation and maintenance

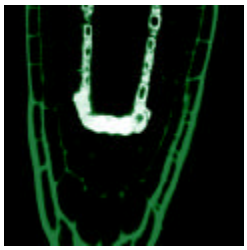
4) Development of Community and Human Resources

The *Arabidopsis* community has developed into an excellent training ground for plant scientists. The changing paradigm of functional genomics will require new types of training to encourage and facilitate lateral, interdisciplinary approaches to problem solving.

Some of the technologies that will be used in the new era of functional genomics research will be beyond the scope of individual labs and some will require sets of biological reagents that are not feasible for individual labs to produce, such as a complete cDNA library and complete protein and metabolite inventories. Instead, a new paradigm will arise in which Genome Technology Centers will serve the research community at large by providing services and by producing new tools using economies of scale. The Centers will be dedicated to the creation of and providing access to genome-wide tools, rather than the application of genome-wide tools to solving specific research problems. The Centers will thereby enable and facilitate the continued participation of individual labs in functional genomic research.



Depending on the status of development and implementation of the technology, such Centers may be financially supported or may operate self-sustained through user fees. In creating genome-wide tools, the Centers must complement



and significantly enable investigators throughout the world. Individual investigators will be at once the main clientele for the Centers and, as the experts in specific biological topics, the dispersed creators of knowledge. The value of this project therefore depends on significant support being available for individual research laboratories throughout the plant biology research community to leverage investment in both the *Arabidopsis* genome sequencing project and the proposed Centers to solve a wide range of specific biological problems.

The structure of Genome Technology Centers, providing services and economies of scale for systems-based data generation, is not consistent with the traditional training of doctoral and post-doctoral researchers, and the traditional output measurement of publications. Therefore, skilled technical assistants and research personnel will be needed. We will also still need traditionally trained doctoral and post-doctoral researchers with skills in plant molecular biology, genetics and biochemistry.

Short to Mid-term goals:

- Support the establishment and maintenance of Genome Technology Centers, including support for technology development and for individual labs seeking the services of the Centers
- Establish summer courses or other intensive specialized workshops, which are an additional effective means for continuing education of established investigators in a rapidly moving field like plant biology and for initiation of non-plant biologists into the world of plant biology

On-going goals:

- Encourage interdisciplinary training that specifically seeks a systems-based approach for both undergraduate and graduate level students
- Support short- and long-term post-doctoral fellowships with a focus on exchange visits

5) International Cooperation: The Multinational Arabidopsis Steering Committee

At the outset of the Multinational Coordinated *Arabidopsis thaliana* Genome Research Project, an *ad hoc* committee was formed, made up of representatives of countries and programs around the world involved in *Arabidopsis* research. Among the goals of this committee were forging relationships and fostering communication among the involved groups. After some time had passed, the committee coalesced into the Multinational Science Steering Committee, and it was members of this committee who ensured communication among the *Arabidopsis* research community at large.

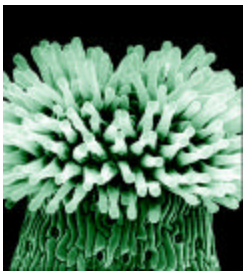
Arabidopsis research has now progressed to a stage in which functional genomics will take the forefront. Participants in this exciting new world of genomic research realized that the same type of communication and coordination provided by the Multinational Science Steering Committee during the first 10 years of the Project will be necessary in the new era. Accordingly, the committee has been renewed under the title Multinational *Arabidopsis* Steering Committee to coordinate various functional genomics activities world-wide.

The Multinational *Arabidopsis* Steering Committee (MASC) will be composed of representatives from each country with major *Arabidopsis* functional genomics efforts or coalition of countries with smaller programs. It is open to any country interested in participating. Selection of MASC representatives is left to the discretion of each country. It will meet once a year in conjunction with the International Conference on *Arabidopsis* Research. Specific responsibilities of the committee are:

- To coordinate programmatic aspects of the *Arabidopsis* research world-wide
- To facilitate open communication and free exchange of data, materials and ideas among the *Arabidopsis* research community
- To monitor and summarize progress of scientific activities of participating laboratories
- To identify needs and opportunities of the *Arabidopsis* research community and communicate them to funding agencies of participating nations
- To periodically update and adjust the course of the Project

Short-term goals:

- Appoint a full time coordinator for the Multinational *Arabidopsis* Steering Committee
- Publish a long-range plan for the Multinational *Arabidopsis* Functional Genomics Project
- Establish and maintain an internet site devoted to functional genomics efforts and communication among members of the world-wide *Arabidopsis* research community



On-going goals:

- Continue to foster international collaboration and coordination of the project
- Continue to monitor progress, periodically reassess the status of the project, and adjust the goals as needed
- Publish periodic progress reports